Kernel Methods for Causal Effect Estimation

Arthur Gretton

Gatsby Computational Neuroscience Unit,
Google DeepMind

Causality and Machine Learning, INI 2026

Outline

Introduction to kernel ridge regression

Causal effect estimation, observed covariates:

- Dose-response curve (DR), heterogeneous response curve (HR), average treatment on treated (ATT), mediation effect, dynamic treatment effect.

Causal effect estimation, hidden covariates:

- ... instrumental variables (and proxy variables)

What's new? What is it good for?

- Treatment A, covariates X, etc can be multivariate, complicated...
- ...by using kernel or adaptive neural net feature representations

Regression assumption: linear functions of features

All learned functions will take the form:

$$\gamma(x) = \gamma^\top \varphi_\theta(x)$$
$$\overset{\text{or}}{=} \langle \gamma, \varphi(x) \rangle_{\mathcal{H}_\mathcal{X}}$$

Regression assumption: linear functions of features

All learned functions will take the form:

$$\gamma(x) = \gamma^\top \varphi_\theta(x)$$
$$\overset{\text{or}}{=} \langle \gamma, \varphi(x) \rangle_{\mathcal{H}_\mathcal{X}}$$

Option 1: Finite dictionaries of learned neural net features $\varphi_\theta(x)$ (linear final layer γ)

Xu, G., A Neural mean embedding approach for back-door and front-door adjustment. (ICLR 23)

Xu, Chen, Srinivasan, de Freitas, Doucet, G. Learning Deep Features in Instrumental Variable Regression. (ICLR 21)

Option 2: Infinite dictionaries of fixed kernel features:

$$\langle \varphi(x_i), \varphi(x) \rangle_{\mathcal{H}_\mathcal{X}} = k(x_i, x)$$

Kernel is feature dot product.

Singh, Xu, G. Kernel Methods for Causal Functions: Dose, Heterogeneous, and Incremental Response Curves. (Biometrika 23)

Singh, Sahani, G. Kernel Instrumental Variable Regression. (NeurIPS 19)

Model fitting: kernel ridge regression

Learn $\gamma_0(x) := \mathbb{E}[\,Y|X = x]$ from features $\varphi(x_i)$ with outcomes y_i:

$$\hat{\gamma} \quad = \quad \arg\min_{\gamma \in \mathcal{H}} \left(\sum_{i=1}^{n} \left(y_i - \langle \gamma, \varphi(x_i) \rangle_{\mathcal{H}_\mathcal{X}} \right)^2 + \lambda \|\gamma\|^2_{\mathcal{H}_\mathcal{X}} \right).$$

Model fitting: kernel ridge regression

Learn $\gamma_0(x) := \mathbb{E}[\,Y|X = x]$ from features $\varphi(x_i)$ with outcomes y_i:

$$\hat{\gamma} \;=\; \arg\min_{\gamma\in\mathcal{H}} \left(\sum_{i=1}^{n} \left(y_i - \langle \gamma, \varphi(x_i)\rangle_{\mathcal{H}_\mathcal{X}} \right)^2 + \lambda \|\gamma\|^2_{\mathcal{H}_\mathcal{X}} \right).$$

Infinite dimensional solution at x:

$$\hat{\gamma}(x) = \widehat{C}_{YX}(\widehat{C}_X + \lambda)^{-1}\varphi(x)$$

$$\widehat{C}_{YX} = \frac{1}{n}\sum_{i=1}^{n}[y_i\,\varphi(x_i)]$$

$$\widehat{C}_X = \frac{1}{n}\sum_{i=1}^{n}[\varphi(x_i)\otimes\varphi(x_i)]$$

Notation : $(f\otimes g)h = f\langle g, h\rangle_{\mathcal{H}_\mathcal{X}}$

Model fitting: kernel ridge regression

Learn $\gamma_0(x) := \mathbb{E}[\,Y|X = x]$ from features $\varphi(x_i)$ with outcomes y_i:

$$\hat{\gamma} \quad = \quad \arg\min_{\gamma \in \mathcal{H}} \left(\sum_{i=1}^{n} \left(y_i - \langle \gamma, \varphi(x_i) \rangle_{\mathcal{H}_\mathcal{X}} \right)^2 + \lambda \|\gamma\|^2_{\mathcal{H}_\mathcal{X}} \right).$$

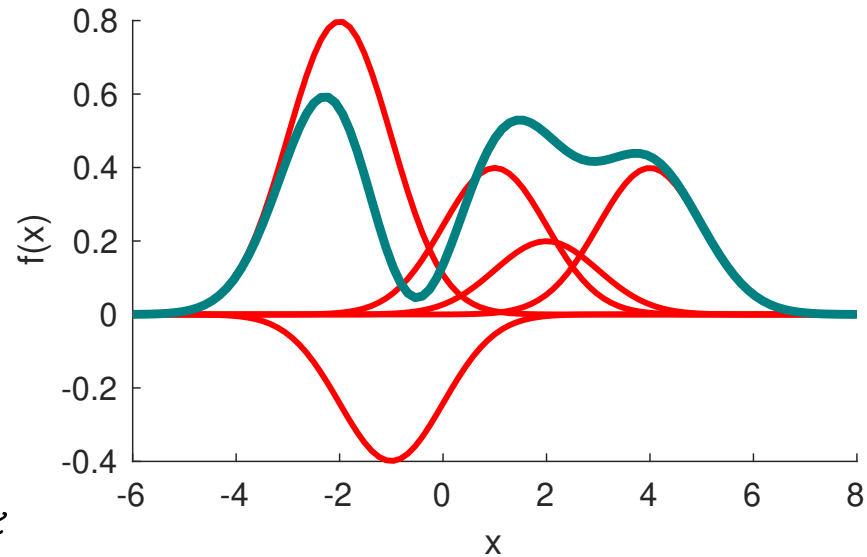
Kernel solution at x
(as weighted sum of y)

$$\hat{\gamma}(x) = \sum_{i=1}^{n} y_i \beta_i(x)$$

$$\beta(x) = (K_{XX} + \lambda I)^{-1} k_{Xx}$$

$$(K_{XX})_{ij} = k(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle_{\mathcal{H}_\mathcal{X}}$$

$$(k_{Xx})_i = k(x_i, x)$$



KRR: consistency in RKHS norm

Define population feature covariance:

$$C_X := \mathbb{E}\left[\varphi(X) \otimes \varphi(X)\right] \quad \langle f, C_X\, g\rangle_{\mathcal{H}_\mathcal{X}} = \mathbb{E}\left[f(X) g(X)\right]$$

$$C_X = \sum_{i=1}^{\infty} \eta_i \left(\sqrt{\eta_i}\, e_i\right) \otimes \left(\sqrt{\eta_i}\, e_i\right)$$

Assume for simplicity: $\operatorname{supp}(P_X) = \mathcal{X}$.

Assume problem well specified:

KRR: consistency in RKHS norm

Define population feature covariance:

$$C_X := \mathbb{E}\left[\varphi(X) \otimes \varphi(X)\right] \quad \langle f, C_X\, g\rangle_{\mathcal{H}_\mathcal{X}} = \mathbb{E}\left[f(X) g(X)\right]$$

$$C_X = \sum_{i=1}^{\infty} \eta_i \left(\sqrt{\eta_i}\, e_i\right) \otimes \left(\sqrt{\eta_i}\, e_i\right)$$

Assume for simplicity: $\operatorname{supp}(P_X) = \mathcal{X}$.

Assume problem well specified:

- $\gamma_* \in \mathcal{H}_\mathcal{X}^c$ where $\mathcal{H}_\mathcal{X}^c \subset \mathcal{H}_\mathcal{X}$, $\quad c \in (1, 2]$

KRR: consistency in RKHS norm

Define population feature covariance:

$$C_X := \mathbb{E}\left[\varphi(X) \otimes \varphi(X)\right] \quad \langle f, C_X\, g\rangle_{\mathcal{H}_\mathcal{X}} = \mathbb{E}\left[f(X) g(X)\right]$$

$$C_X = \sum_{i=1}^{\infty} \eta_i \left(\sqrt{\eta_i}\, e_i\right) \otimes \left(\sqrt{\eta_i}\, e_i\right)$$

Assume for simplicity: $\operatorname{supp}(P_X) = \mathcal{X}$.

Assume problem well specified:

- $\gamma_* \in \mathcal{H}_\mathcal{X}^c$ where $\mathcal{H}_\mathcal{X}^c \subset \mathcal{H}_\mathcal{X}, \quad c \in (1, 2]$

$$\gamma_* = \sum_i \gamma_{*,i}\, e_i \quad \text{where} \quad \|\gamma_*\|^2_{\mathcal{H}_\mathcal{X}^c} = \sum_{i=1}^{\infty} \frac{\gamma_{*,i}^2}{\eta_i^c} < \infty$$

  - Larger $c \implies$ smoother $\gamma_* \implies$ easier problem.

KRR: consistency in RKHS norm

Define population feature covariance:

$$C_X := \mathbb{E}\left[\varphi(X) \otimes \varphi(X)\right] \quad \langle f, C_X\, g\rangle_{\mathcal{H}_\mathcal{X}} = \mathbb{E}\left[f(X) g(X)\right]$$

$$C_X = \sum_{i=1}^{\infty} \eta_i \left(\sqrt{\eta_i}\, e_i\right) \otimes \left(\sqrt{\eta_i}\, e_i\right)$$

Assume for simplicity: $\operatorname{supp}(P_X) = \mathcal{X}$.

Assume problem well specified:

- $\gamma_* \in \mathcal{H}_\mathcal{X}^c$ where $\mathcal{H}_\mathcal{X}^c \subset \mathcal{H}_\mathcal{X}$, $\quad c \in (1, 2]$

$$\gamma_* = \sum_i \gamma_{*,i}\, e_i \quad \text{where} \quad \|\gamma_*\|^2_{\mathcal{H}_\mathcal{X}^c} = \sum_{i=1}^{\infty} \frac{\gamma_{*,i}^2}{\eta_i^c} < \infty$$

  - Larger $c \implies$ smoother $\gamma_* \implies$ easier problem.
- Eigenspectrum decay of input feature covariance, $\eta_j \sim j^{-b}$, $b \geq 1$
  - Larger $b \implies$ easier problem

KRR: consistency in RKHS norm

Consistency for models in class $\mathcal{M}(c, b)$ [A, Theorem 1.ii]

$$\|\hat{\gamma} - \gamma_*\|^2_{L_2(\mathcal{X})} = O_P\left(n^{-\frac{c}{c+1/b}}\right) \qquad \|\hat{\gamma} - \gamma_*\|^2_{\mathcal{H}_\mathcal{X}} = O_P\left(n^{-\frac{c-1}{c+1/b}}\right)$$

Example:

$\mathcal{X} = \mathbb{R}^d$, $\mathcal{H}_\mathcal{X} := W^{\nu,2}(\mathcal{X})$ Sobolev space with $\nu > \frac{d}{2}$ (Matérn kernel) gives $b = \frac{2\nu}{d}$ [B, Corollary 2].

Then $\gamma_* \in \mathcal{H}^c_\mathcal{X}$ with $c \in (1, 2]$ gives consistency in RKHS and L_2.

[A] Fischer, Steinwart (2020). Sobolev norm learning rates for regularized least-squares algorithms.
[B] Li, Meunier, Mollenhauer, G (2024), Towards Optimal Sobolev Norm Rates for VV RLS Algorithm

Causal effects, observed covariates

Kernels (Biometrika 2023):

arXiv > econ > arXiv:2010.04855 Search... Help | Advan

Economics > Econometrics

[Submitted on 10 Oct 2020 (v1), last revised 23 Aug 2022 (this version, v6)]

Kernel Methods for Causal Functions: Dose, Heterogeneous, and Incremental Response Curves

Rahul Singh, Liyuan Xu, Arthur Gretton

NN features (ICLR 2023):

arXiv > cs > arXiv:2210.06610 Search... Help | Advan

Computer Science > Machine Learning

[Submitted on 12 Oct 2022]

A Neural Mean Embedding Approach for Back-door and Front-door Adjustment

Liyuan Xu, Arthur Gretton

Code for NN and kernel causal estimation with observed covariates:

https://github.com/liyuan9988/DeepFrontBackDoor/

Causal effects, observed covariates

Kernel features (Biometrika 2023):

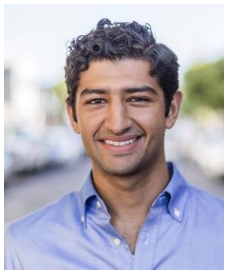
arXiv > econ > arXiv:2010.04855

Economics > Econometrics

[Submitted on 10 Oct 2020 (v1), last revised 23 Aug 2022 (this version, v6)]

Kernel Methods for Causal Functions: Dose, Heterogeneous, and Incremental Response Curves

Rahul Singh, Liyuan Xu, Arthur Gretton



NN features (ICLR 2023):

arXiv > cs > arXiv:2210.06610

Computer Science > Machine Learning

[Submitted on 12 Oct 2022]

A Neural Mean Embedding Approach for Back-door and Front-door Adjustment

Liyuan Xu, Arthur Gretton

Code for NN and kernel causal estimation with observed covariates:

https://github.com/liyuan9988/DeepFrontBackDoor/

Dose-response curve

Potential outcome (intervention):

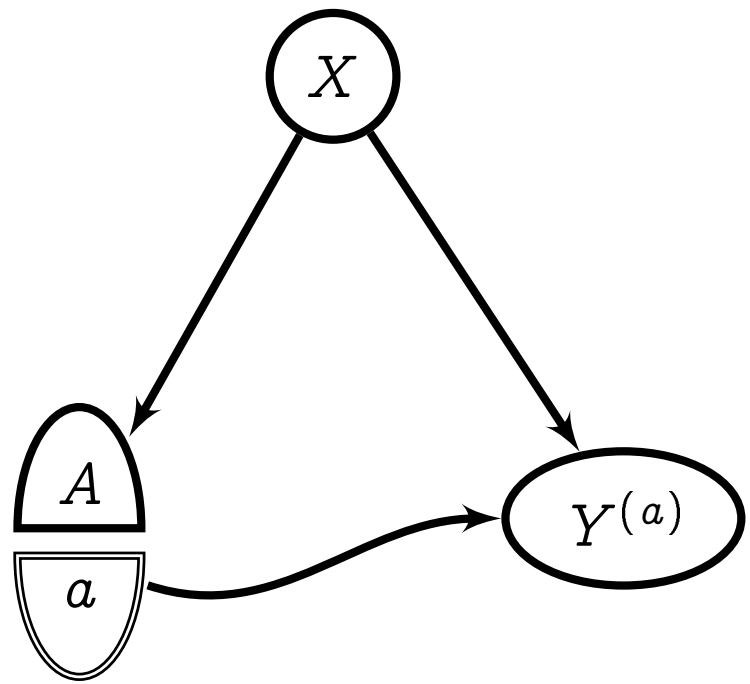
$$\mathrm{DR}(a) := \mathbb{E}[Y^{(a)}] = \int \mathbb{E}[Y|a,x]\,dp(x)$$

(the average structural function; for continuous a, the dose-response curve).

Assume: (1) Stable Unit Treatment Value Assumption (aka "no interference"), (2) Conditional exchangeability $Y^{(a)} \perp\!\!\!\perp A|X$. (3) Overlap.

Example: US job corps, training for disadvantaged youths:

- A: treatment (training hours)
- Y: outcome (percentage employment)
- X: covariates (age, education, marital status, ...)

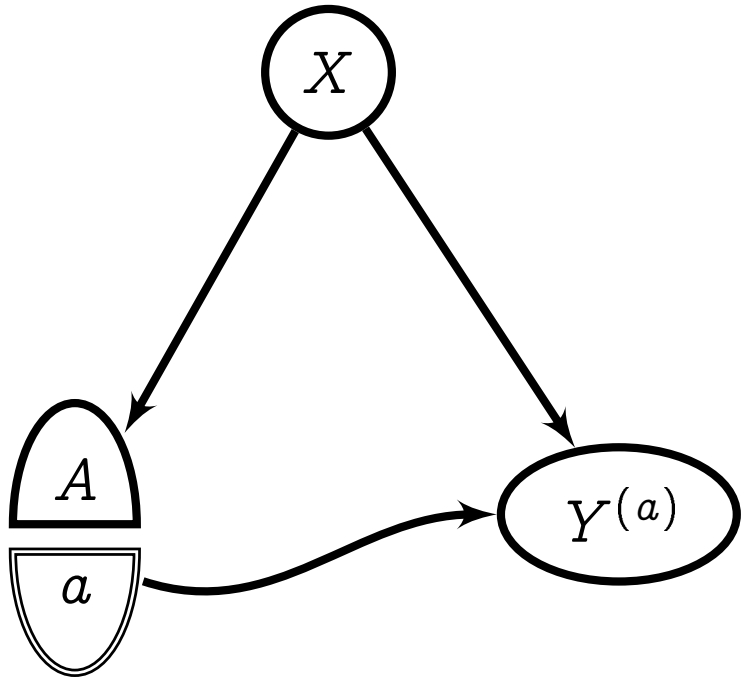


Multiple inputs via products of kernels

The expected output is a function of two arguments

$$\mathbb{E}[\,Y \,|\, a, x\,] = \langle \gamma_*, \varphi(a) \otimes \varphi(x) \rangle_{\mathcal{H}_\mathcal{A} \otimes \mathcal{H}_\mathcal{X}}$$

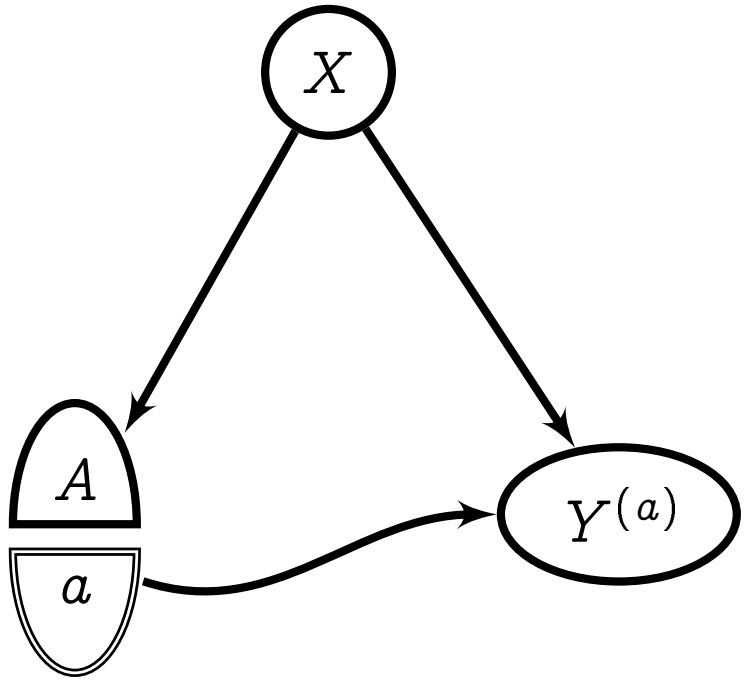
(argument of kernel/feature map indicates feature space)



Multiple inputs via products of kernels

The expected output is a function of two arguments

$$\mathbb{E}[\,Y \,|\, a, x\,] = \langle \gamma_*, \varphi(a) \otimes \varphi(x) \rangle_{\mathcal{H}_\mathcal{A} \otimes \mathcal{H}_\mathcal{X}}$$



(argument of kernel/feature map indicates feature space)

Ridge regression solution:

$$\hat{\gamma}(x, a) = \sum_{i=1}^{n} y_i \beta_i(a, x), \quad \beta(a, x) = \left[K_{AA} \odot K_{XX} + \lambda I\right]^{-1} K_{Aa} \odot K_{Xx}$$

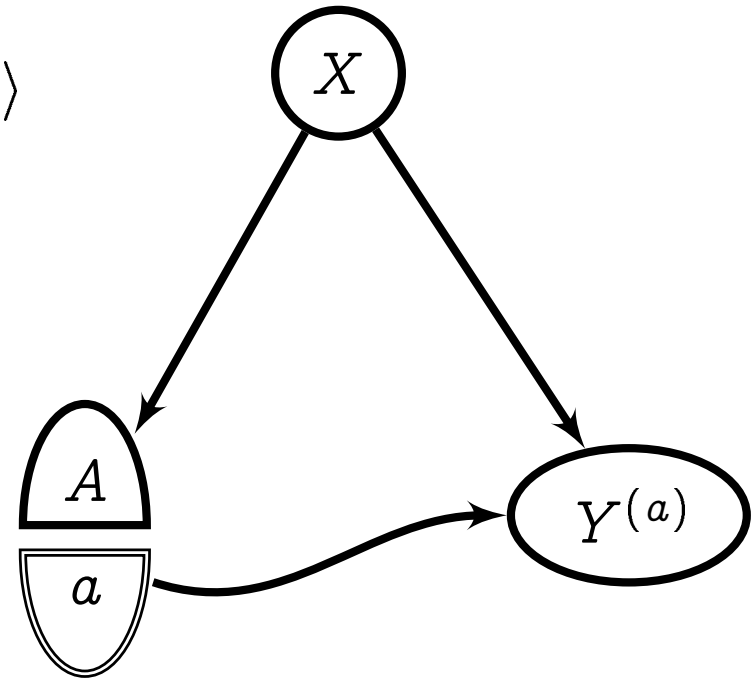
Dose-response curve

Well-specified setting:

$$\mathbb{E}[\,Y\,|\,a,x] =: \gamma_*(a,x) = \langle \gamma_*, \varphi(a) \otimes \varphi(x) \rangle$$

DR as feature space dot product:

$$\begin{aligned} \mathrm{DR}(a) &= \mathbb{E}[\gamma_*(a, X)] \\ &= \mathbb{E}\left[\langle \gamma_*, \varphi(a) \otimes \varphi(X) \rangle\right] \end{aligned}$$



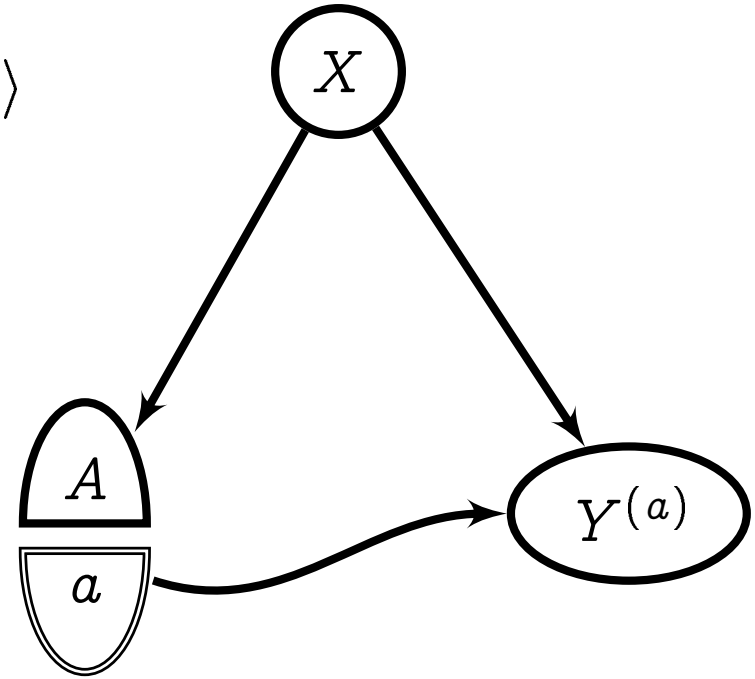
Dose-response curve

Well-specified setting:

$$\mathbb{E}[\,Y\,|\,a,x\,] =: \gamma_*(a,x) = \langle \gamma_*, \varphi(a) \otimes \varphi(x) \rangle$$

DR as feature space dot product:

$$\begin{aligned} \mathrm{DR}(a) &= \mathbb{E}[\gamma_*(a,X)] \\ &= \mathbb{E}\left[\langle \gamma_*, \varphi(a) \otimes \varphi(X) \rangle\right] \\ &= \langle \gamma_*, \varphi(a) \otimes \underbrace{\mu_X}_{\mathbb{E}[\varphi(X)]} \rangle \end{aligned}$$



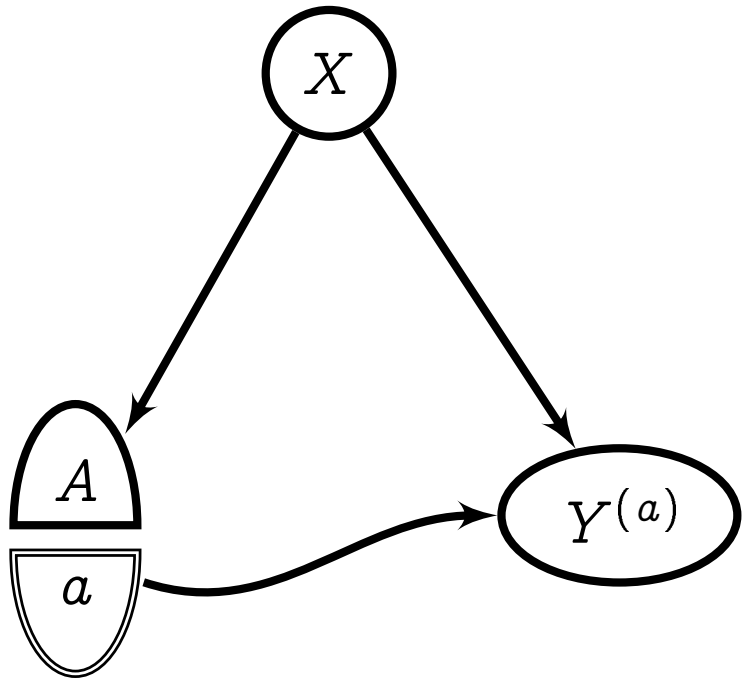
Feature map of probability $P(X)$,

$$\mu_X = [\ldots \mathbb{E}\left[\varphi_i(X)\right] \ldots]$$

DR: example

US job corps: training for disadvantaged youths:

- X: covariate/context (age, education, marital status, ...)
- A: treatment (training hours)
- Y: outcome (percent employment)



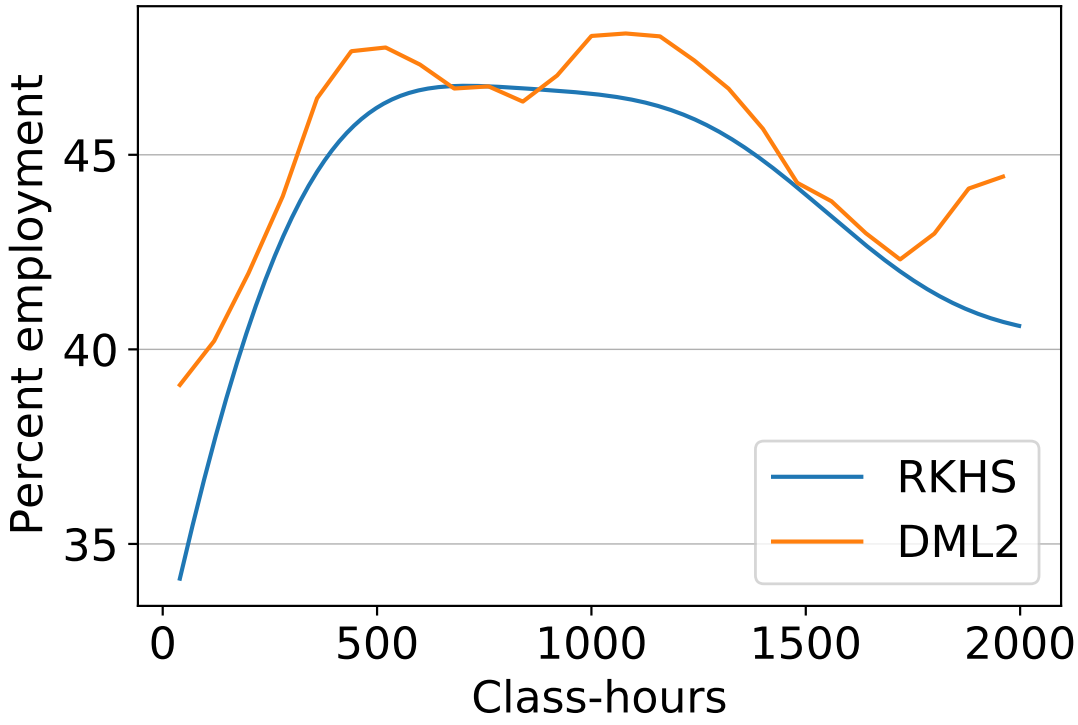
Empirical DR:

$$\widehat{\mathrm{DR}}(a) = \widehat{\mathbb{E}}\left[\langle \hat{\gamma}, \varphi(a) \otimes \varphi(X) \rangle\right]$$
$$= \frac{1}{n} \sum_{i=1}^{n} Y^\top (K_{AA} \odot K_{XX} + n\lambda I)^{-1} (K_{Aa} \odot K_{Xx_i})$$

Schochet, Burghardt, and McConnell (2008). Does Job Corps work? Impact findings from the national Job Corps study.

Singh, Xu, G (2023).

DR: results



- First 12.5 weeks of classes confer employment gain: from 35% to 47%.
- [RKHS] is our $\widehat{\mathrm{DR}}(a)$.
- [DML2] Colangelo, Lee (2020), Double debiased machine learning nonparametric inference with continuous treatments.

Singh, Xu, G (2023)

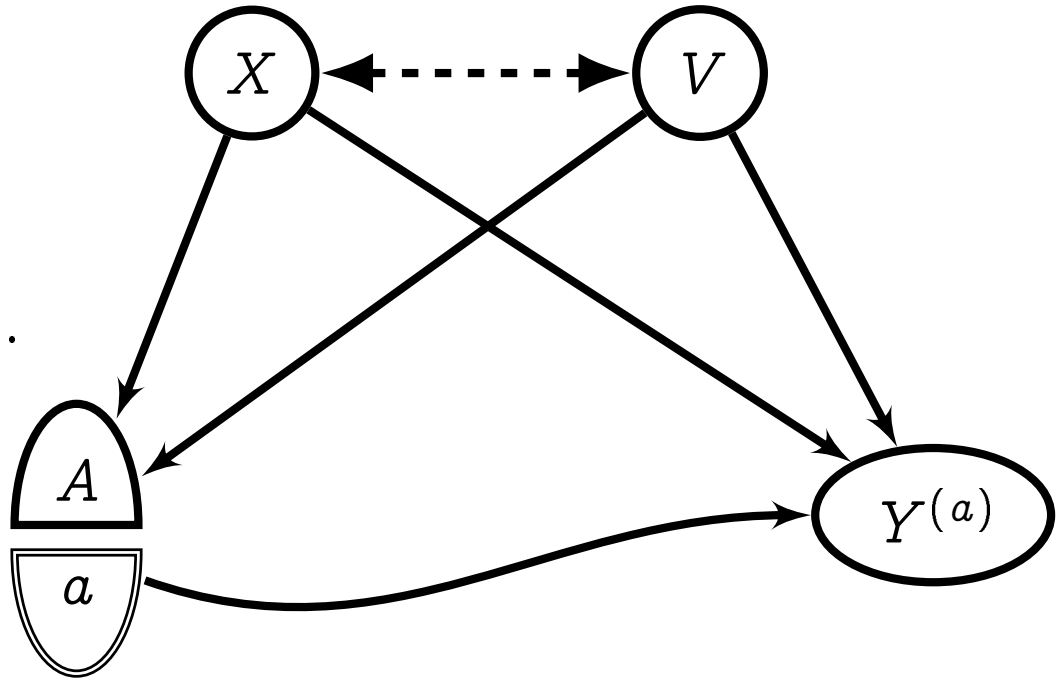
Heterogeneous response curve

Well-specified setting:

$$\mathbb{E}[Y|a,x,v] =: \gamma_*(a,x,v)$$
$$= \langle \gamma_*, \varphi(a) \otimes \varphi(x) \otimes \varphi(v) \rangle .$$

Heterogeneous response:

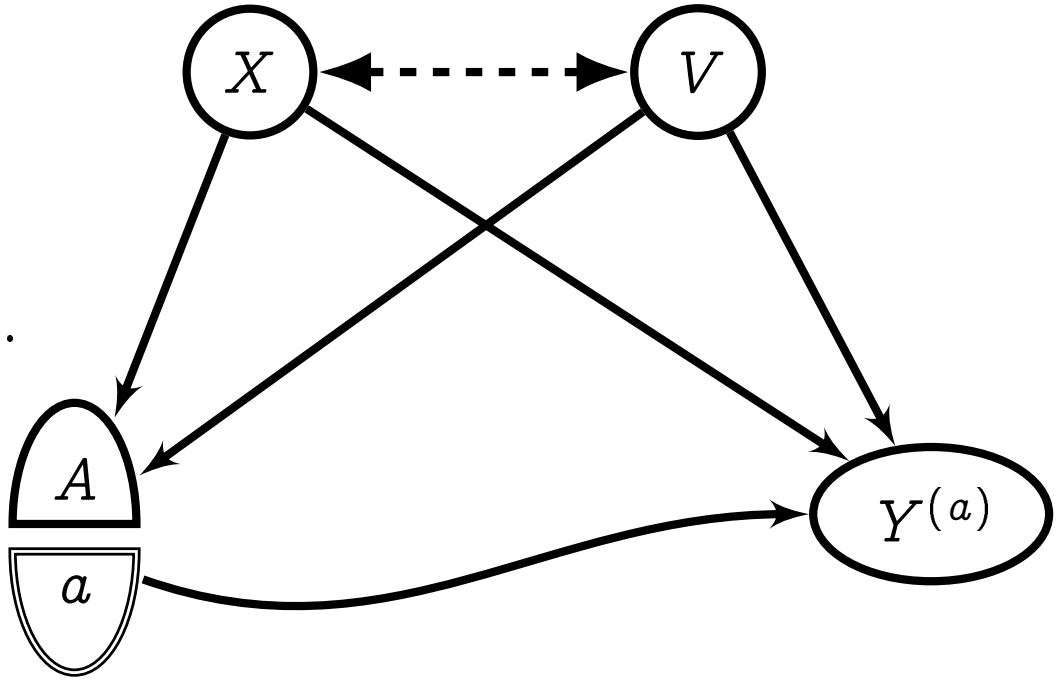
$$\mathrm{HR}(a,v)$$
$$= \mathbb{E}\left[Y^{(a)} | V = v \right]$$



Heterogeneous response curve

Well-specified setting:

$$\mathbb{E}[Y|a,x,v] =: \gamma_*(a,x,v)$$
$$= \langle \gamma_*, \varphi(a) \otimes \varphi(x) \otimes \varphi(v) \rangle .$$



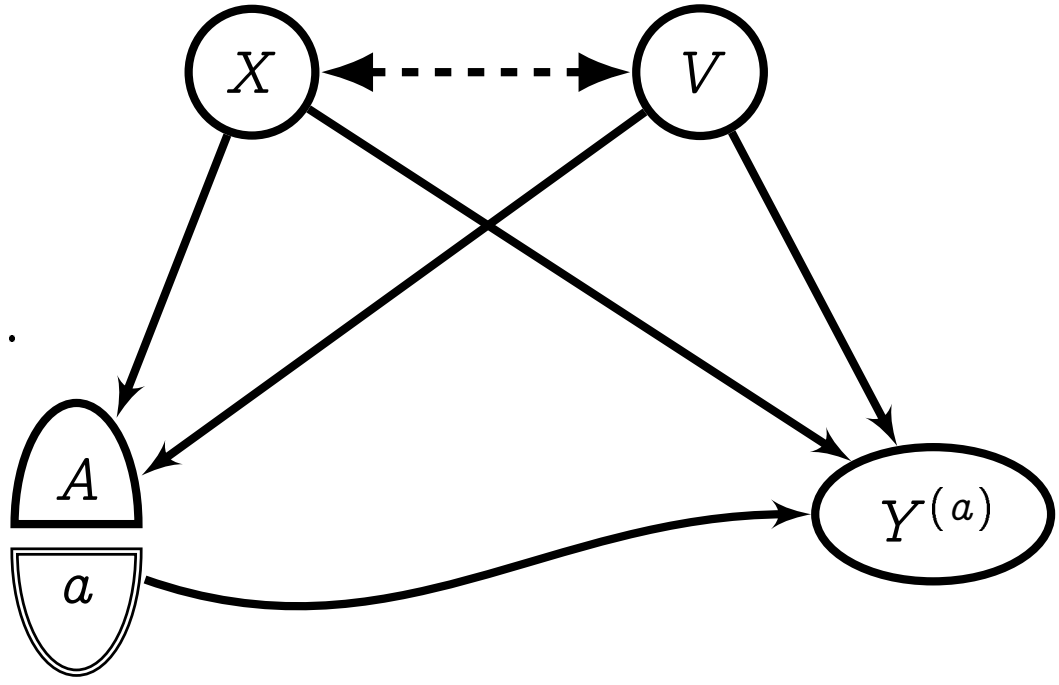
Heterogeneous response:

$$\mathrm{HR}(a,v)$$
$$= \mathbb{E}\left[Y^{(a)} | V = v \right]$$
$$= \mathbb{E}\left[\langle \gamma_*, \varphi(a) \otimes \varphi(X) \otimes \varphi(V) \rangle \,|\, V = v \right]$$

Heterogeneous response curve

Well-specified setting:

$$\mathbb{E}[Y|a,x,v] =: \gamma_*(a,x,v)$$
$$= \langle \gamma_*, \varphi(a) \otimes \varphi(x) \otimes \varphi(v) \rangle .$$



Heterogeneous response:

$$\mathrm{HR}(a,v)$$
$$= \mathbb{E}\left[Y^{(a)} | V = v \right]$$
$$= \mathbb{E}\left[\langle \gamma_*, \varphi(a) \otimes \varphi(X) \otimes \varphi(V) \rangle \,|\, V = v \right]$$
$$= \ldots ?$$

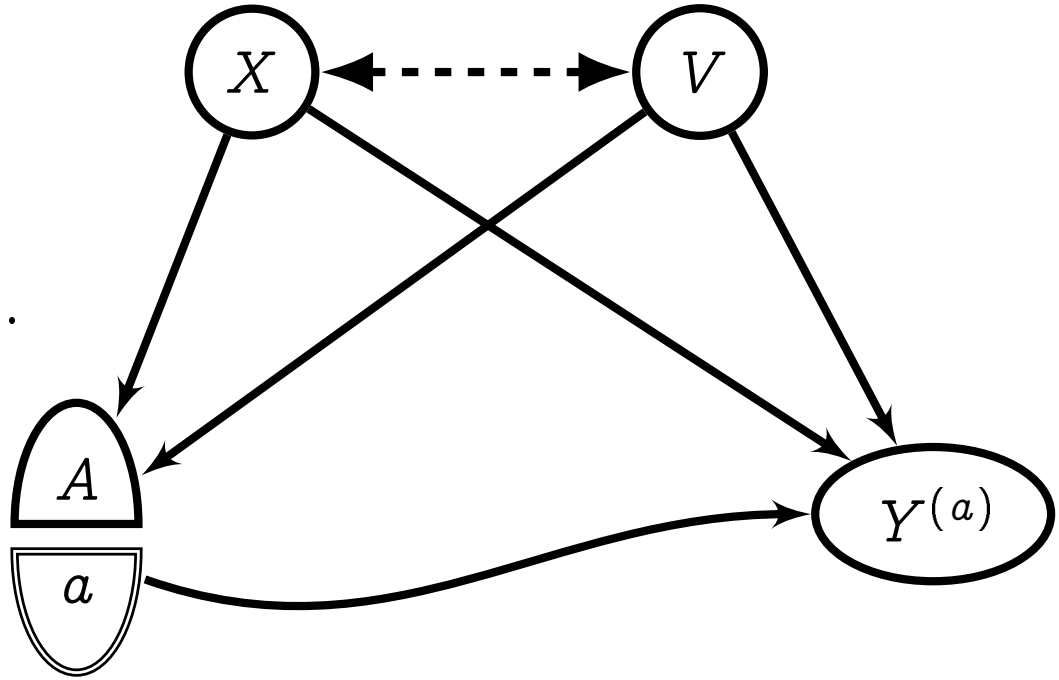
How to take conditional expectation?

Density estimation for $p(X|V=v)$? Sample from $p(X|V=v)$?

Heterogeneous response curve

Well-specified setting:

$$\mathbb{E}[Y|a,x,v] =: \gamma_*(a,x,v)$$
$$= \langle \gamma_*, \varphi(a) \otimes \varphi(x) \otimes \varphi(v) \rangle .$$



Heterogeneous response:

$$\mathrm{HR}(a,v)$$
$$= \mathbb{E}\left[Y^{(a)} | V = v\right]$$
$$= \mathbb{E}\left[\langle \gamma_*, \varphi(a) \otimes \varphi(X) \otimes \varphi(V) \rangle | V = v\right]$$
$$= \langle \gamma_*, \varphi(a) \otimes \underbrace{\mathbb{E}[\varphi(X)|V=v]}_{\mu_{X|V=v}} \otimes \varphi(v) \rangle$$

Learn conditional mean embedding: $\mu_{X|V=v} := \mathbb{E}_X[\varphi(X)|V=v]$

Regressing from feature space to feature space

Our goal: an operator $F_* : \mathcal{H}_\mathcal{V} \to \mathcal{H}_\mathcal{X}$ such that

$$F_* \varphi(v) = \mathbb{E}_X \left[\varphi(X) \middle| V = v \right] =: \mu_{X|V=v}$$

Song, Huang, Smola, Fukumizu (2009). Hilbert space embeddings of conditional distributions with applications to dynamical systems.

Grunewalder, Lever, Baldassarre, Patterson, G, Pontil (2012). Conditional mean embeddings as regressors.

Grunewalder, G, Shawe-Taylor (2013) Smooth operators.

Li, Meunier, Mollenhauer, G (2022), Optimal Rates for Regularized Conditional Mean Embedding Learning

Regressing from feature space to feature space

Our goal: an operator $F_* : \mathcal{H}_\mathcal{V} \to \mathcal{H}_\mathcal{X}$ such that

$$F_* \varphi(v) = \mathbb{E}_X [\varphi(X) | V = v] =: \mu_{X|V=v}$$

Assume F_* Hilbert-Schmidt,

$$F_* \in S_2(\mathcal{H}_\mathcal{V}, \mathcal{H}_\mathcal{X})$$

Implied smoothness assumption:

$$\mathbb{E}[h(X) | V = v] \in \mathcal{H}_\mathcal{V} \quad \forall h \in \mathcal{H}_\mathcal{X}$$

Song, Huang, Smola, Fukumizu (2009). Hilbert space embeddings of conditional distributions with applications to dynamical systems.

Grunewalder, Lever, Baldassarre, Patterson, G, Pontil (2012). Conditional mean embeddings as regressors.

Grunewalder, G, Shawe-Taylor (2013) Smooth operators.

Li, Meunier, Mollenhauer, G (2022), Optimal Rates for Regularized Conditional Mean Embedding Learning

Regressing from feature space to feature space

Our goal: an operator $F_* : \mathcal{H}_\mathcal{V} \to \mathcal{H}_\mathcal{X}$ such that

$$F_* \varphi(v) = \mathbb{E}_X [\varphi(X) | V = v] =: \mu_{X|V=v}$$

Assume F_* Hilbert-Schmidt,

$$F_* \in S_2(\mathcal{H}_\mathcal{V}, \mathcal{H}_\mathcal{X})$$

Implied smoothness assumption:

$$\mathbb{E}[h(X) | V = v] \in \mathcal{H}_\mathcal{V} \quad \forall h \in \mathcal{H}_\mathcal{X}$$

A Smooth Operator

Song, Huang, Smola, Fukumizu (2009). Hilbert space embeddings of conditional distributions with applications to dynamical systems.

Grunewalder, Lever, Baldassarre, Patterson, G, Pontil (2012). Conditional mean embeddings as regressors.

Grunewalder, G, Shawe-Taylor (2013) Smooth operators.

Li, Meunier, Mollenhauer, G (2022), Optimal Rates for Regularized Conditional Mean Embedding Learning

Regressing from feature space to feature space

Our goal: an operator $F_* : \mathcal{H}_\mathcal{V} \to \mathcal{H}_\mathcal{X}$ such that

$$F_* \varphi(v) = \mathbb{E}_X [\varphi(X) | V = v] =: \mu_{X|V=v}$$

Assume F_* Hilbert-Schmidt,

$$F_* \in S_2(\mathcal{H}_\mathcal{V}, \mathcal{H}_\mathcal{X})$$

Implied smoothness assumption:

$$\mathbb{E}[h(X) | V = v] \in \mathcal{H}_\mathcal{V} \quad \forall h \in \mathcal{H}_\mathcal{X}$$

Kernel ridge regression from $\varphi(v)$ to infinite features $\varphi(x)$:

$$\widehat{F} = \operatorname*{argmin}_{F \in S_2(\mathcal{H}_\mathcal{V}, \mathcal{H}_\mathcal{X})} \sum_{\ell=1}^{n} \|\varphi(x_\ell) - F\varphi(v_\ell)\|^2_{\mathcal{H}_\mathcal{X}} + \lambda_2 \|F\|^2_{HS}$$

Song, Huang, Smola, Fukumizu (2009). Hilbert space embeddings of conditional distributions with applications to dynamical systems.

Grunewalder, Lever, Baldassarre, Patterson, G, Pontil (2012). Conditional mean embeddings as regressors.

Grunewalder, G, Shawe-Taylor (2013) Smooth operators.

Li, Meunier, Mollenhauer, G (2022), Optimal Rates for Regularized Conditional Mean Embedding Learning

Regressing from feature space to feature space

Our goal: an operator $F_* : \mathcal{H}_\mathcal{V} \to \mathcal{H}_\mathcal{X}$ such that

$$F_* \varphi(v) = \mathbb{E}_X [\varphi(X) | V = v] =: \mu_{X|V=v}$$

Assume F_* Hilbert-Schmidt,

$$F_* \in S_2(\mathcal{H}_\mathcal{V}, \mathcal{H}_\mathcal{X})$$

Implied smoothness assumption:

$$\mathbb{E}[h(X) | V = v] \in \mathcal{H}_\mathcal{V} \quad \forall h \in \mathcal{H}_\mathcal{X}$$

Kernel ridge regression from $\varphi(v)$ to infinite features $\varphi(x)$:

$$\widehat{F} = \operatorname*{argmin}_{F \in S_2(\mathcal{H}_\mathcal{V}, \mathcal{H}_\mathcal{X})} \sum_{\ell=1}^{n} \|\varphi(x_\ell) - F\varphi(v_\ell)\|^2_{\mathcal{H}_\mathcal{X}} + \lambda_2 \|F\|^2_{HS}$$

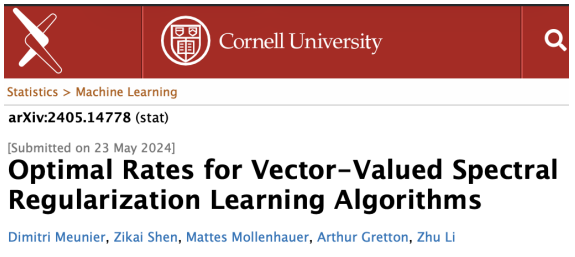
Ridge regression solution:

$$\mu_{X|V=v} := \mathbb{E}[\varphi(X) | V = v] \approx \widehat{F}\varphi(v) = \sum_{\ell=1}^{n} \varphi(x_\ell)\beta_\ell(v)$$

$$\beta(v) = [K_{VV} + \lambda_2 I]^{-1} k_{Vv}$$

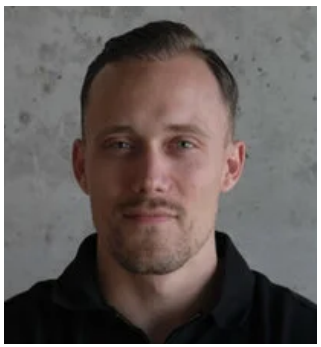
Consistency of Vector Valued Least Squares

NeurIPS (2022)



NeurIPS (2024)

JMLR (2024)



Consistency of Vector Valued Least Squares

Define C_V as covariance of features $\varphi(v)$,

$$C_V = \mathbb{E}(\varphi(V) \otimes \varphi(V)), \qquad \mathbb{E}_V(f(V)g(V)) = \langle f, C_V g \rangle_{\mathcal{H}_\mathcal{V}}$$

- Assume problem well specified [C, Assumption 4]

$$F_* \in S_2(\mathcal{H}_\mathcal{V}^{c_1}, \mathcal{H}_\mathcal{X}), \quad c_1 \in (1, 2],$$

  larger $c_1 \implies$ smoother $F_* \implies$ easier problem.
- Eigenspectrum of C_V decays as $\eta_j(C_V) \sim j^{-b_1}$, $b_1 \geq 1$.

[A] Li, Meunier, Mollenhauer, G (2022), Optimal Rates for Regularized CME Learning
[B] Li, Meunier, Mollenhauer, G (2024), Towards Optimal Sobolev Norm Rates for VV RLS Algorithm
[C] Singh, Xu, G (2023)
Grunewalder, Lever, Baldassarre, Patterson, G, Pontil (2012).
Caponnetto, De Vito (2007).

Consistency of Vector Valued Least Squares

Define C_V as covariance of features $\varphi(v)$,

$$C_V = \mathbb{E}(\varphi(V) \otimes \varphi(V)), \qquad \mathbb{E}_V(f(V)g(V)) = \langle f, C_V g \rangle_{\mathcal{H}_\mathcal{V}}$$

- Assume problem well specified [C, Assumption 4]

$$F_* \in S_2(\mathcal{H}_\mathcal{V}^{c_1}, \mathcal{H}_\mathcal{X}), \quad c_1 \in (1, 2],$$

larger $c_1 \implies$ smoother $F_* \implies$ easier problem.

- Eigenspectrum of C_V decays as $\eta_j(C_V) \sim j^{-b_1}$, $b_1 \geq 1$.

Consistency [A, Theorem 2, Theorem 3] (minimax)

$$\left\| \widehat{F} - F_* \right\|_{\text{HS}}^2 = O_P\left(n^{-\frac{c_1 - 1}{c_1 + 1/b_1}} \right),$$

best rate is $O_P(n^{-1/4})$.

[A] Li, Meunier, Mollenhauer, G (2022), Optimal Rates for Regularized CME Learning

[B] Li, Meunier, Mollenhauer, G (2024), Towards Optimal Sobolev Norm Rates for VV RLS Algorithm

[C] Singh, Xu, G (2023)

Grunewalder, Lever, Baldassarre, Patterson, G, Pontil (2012).

Caponnetto, De Vito (2007).

Heterogeneous Response from Samples

Empirical HR:

$$\widehat{\mathrm{HR}}(a, v) = Y^\top (K_{AA} \odot K_{XX} \odot K_{VV} + n\lambda I)^{-1} (K_{Aa} \odot \underbrace{K_{XX}(K_{VV} + n\lambda_1 I)^{-1} K_{Vv}}_{\text{from } \hat{\mu}_{X \mid V=v}} \odot K_{Vv})$$

[A] Singh, Xu, G. (2023)

Heterogeneous Response from Samples

Empirical HR:

$$\widehat{\mathrm{HR}}(a, v) = Y^\top (K_{AA} \odot K_{XX} \odot K_{VV} + n\lambda I)^{-1} (K_{Aa} \odot \underbrace{K_{XX}(K_{VV} + n\lambda_1 I)^{-1} K_{Vv}}_{\text{from } \hat{\mu}_{X \mid V=v}} \odot K_{Vv})$$

Consistency: [A, Theorem 2]

$$\|\widehat{HR} - \mathrm{HR}\|_\infty = O_P\left(n^{-\frac{1}{2}\frac{c-1}{c+1/b}} + n^{-\frac{1}{2}\frac{c_1-1}{c_1+1/b_1}}\right).$$

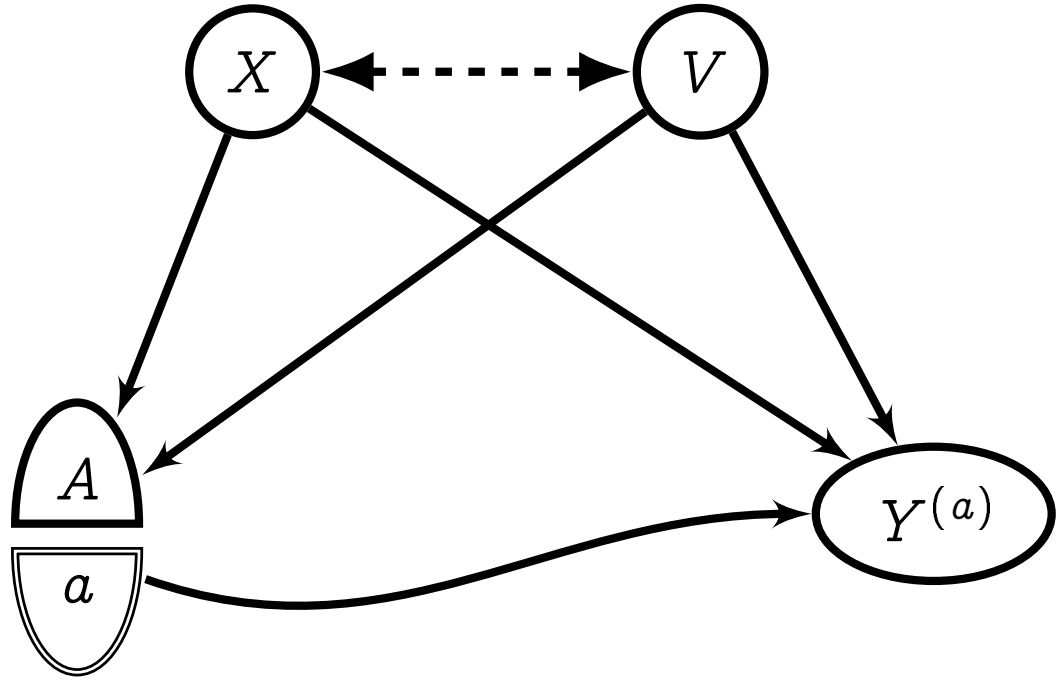
Follows from consistency of $\widehat{F}$ and $\hat{\gamma}$

[A] Singh, Xu, G. (2023)

Heterogeneous response: example

US job corps:

- X: confounder/context (education, marital status, ...)
- A: treatment (training hours)
- Y: outcome (percent employed)
- V: age

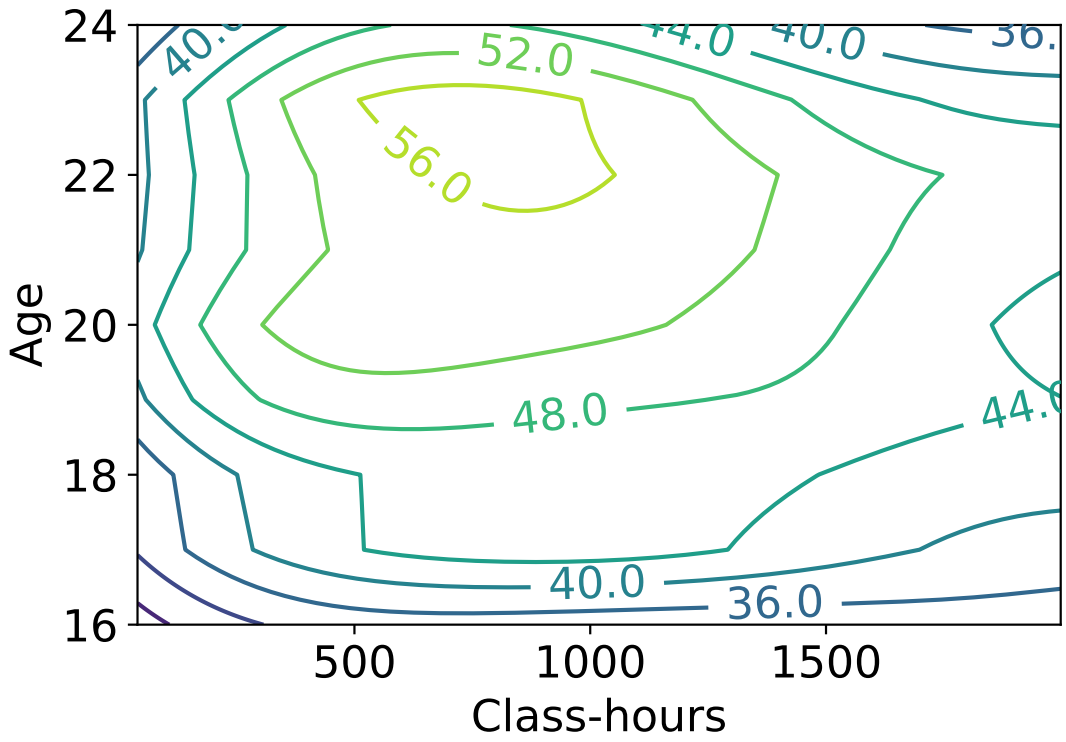


Empirical HR:

$$\widehat{\mathrm{HR}}(a, v) = \langle \hat{\gamma}_0, \varphi(a) \otimes \underbrace{\hat{F}\varphi(v)}_{\widehat{\mathbb{E}}[\varphi(X)\,|\,V=v]} \otimes \varphi(v) \rangle$$

[A] Singh, Xu, G. (2023)

Heterogeneous response: results



Average percentage employment $Y^{(a)}$ for class hours a, conditioned on age v. Given around 12-14 weeks of classes:

- 16 y/o: employment increases from 28% to at most 36%.
- 22 y/o: percent employment increases from 40% to 56%.

[A] Singh, Xu, G. (2023)

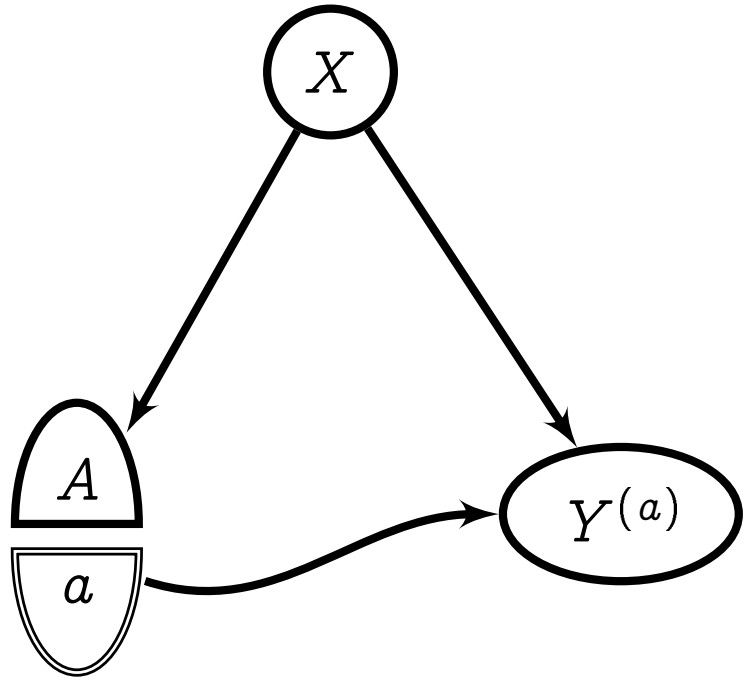
Counterfactual: average treatment on treated

Conditional mean:

$$\mathbb{E}[\,Y\,|\,a, x] = \gamma_*(a, x)$$

Average treatment on treated:

$$\begin{aligned}&\mathrm{ATT}(a, a') \\ &= \mathbb{E}[y^{(a')}|A = a]\end{aligned}$$



Empirical ATT:

$$\widehat{\mathrm{ATT}}(a, a')$$

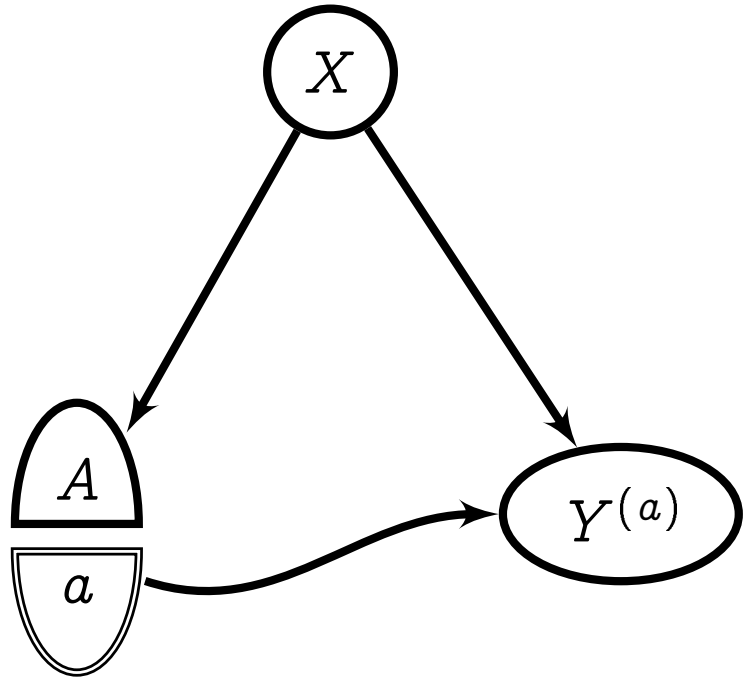
Counterfactual: average treatment on treated

Conditional mean:

$$\mathbb{E}[\,Y\,|\,a, x] = \gamma_*(a, x) = \langle \gamma_0, \varphi(a) \otimes \varphi(x) \rangle$$

Average treatment on treated:

$$\begin{aligned} &\mathrm{ATT}(a, a') \\ &= \mathbb{E}[y^{(a')} | A = a] \end{aligned}$$



Empirical ATT:

$$\widehat{\mathrm{ATT}}(a, a')$$

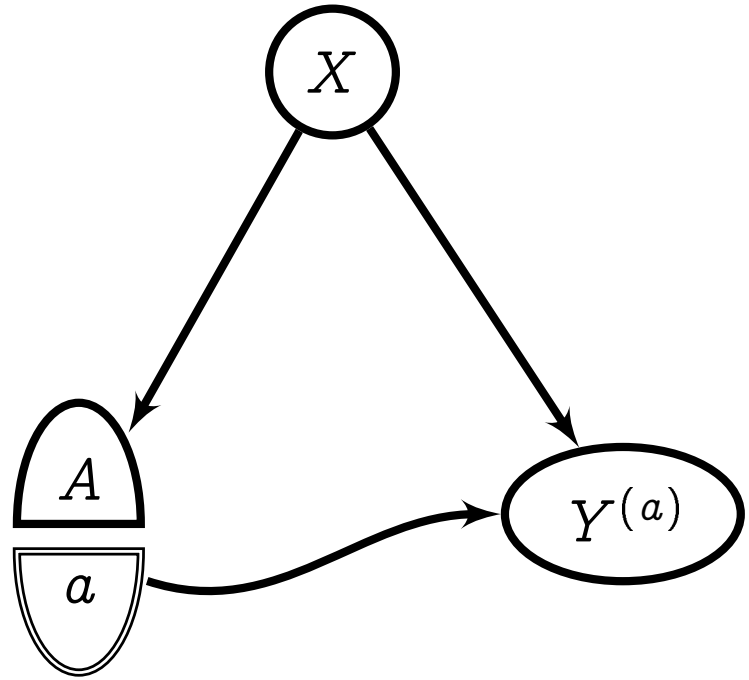
Counterfactual: average treatment on treated

Conditional mean:

$$\mathbb{E}[\,Y\,|\,a, x] = \gamma_*(a, x)$$

Average treatment on treated:

$$\begin{aligned}
&\mathrm{ATT}(a, a') \\
&= \mathbb{E}[y^{(a')}|A = a] \\
&= \mathbb{E}_P\left[\langle \gamma_*, \varphi(a') \otimes \varphi(X)\rangle \,|A = a\right] \\
&= \langle \gamma_*, \varphi(a') \otimes \underbrace{\mathbb{E}_P[\varphi(X)|A = a]}_{\mu_{X|A=a}}\rangle
\end{aligned}$$



Empirical ATT:

$$\widehat{\mathrm{ATT}}(a, a')$$

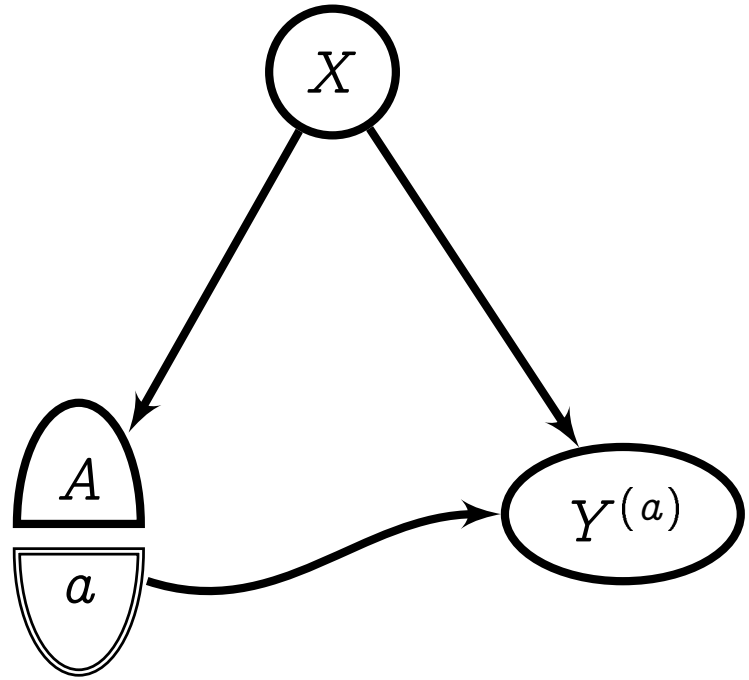
Counterfactual: average treatment on treated

Conditional mean:

$$\mathbb{E}[\,Y\,|\,a, x] = \gamma_*(a, x)$$

Average treatment on treated:

$$\begin{aligned}
&\mathrm{ATT}(a, a') \\
&= \mathbb{E}[y^{(a')}|A = a] \\
&= \mathbb{E}_P\left[\langle \gamma_*, \varphi(a') \otimes \varphi(X)\rangle \,|A = a\right] \\
&= \langle \gamma_*, \varphi(a') \otimes \underbrace{\mathbb{E}_P[\varphi(X)|A = a]}_{\mu_{X|A=a}}\rangle
\end{aligned}$$



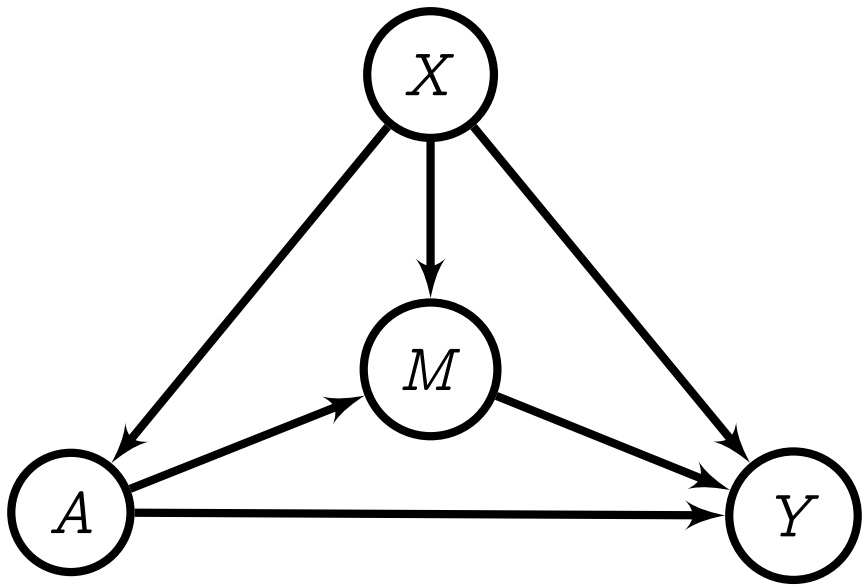
Empirical ATT:

$$\begin{aligned}
&\widehat{\mathrm{ATT}}(a, a') \\
&= Y^\top (K_{AA} \odot K_{XX} + n\lambda I)^{-1} (K_{Aa'} \odot \underbrace{K_{XX}(K_{AA} + n\lambda_1 I)^{-1} K_{Aa}}_{\text{from } \hat{\mu}_{X|A=a}})
\end{aligned}$$

Mediation analysis

- Direct path from treatment A to effect Y
- Indirect path $A \to M \to Y$
- X: context

Is the effect Y mainly due to A? To M?

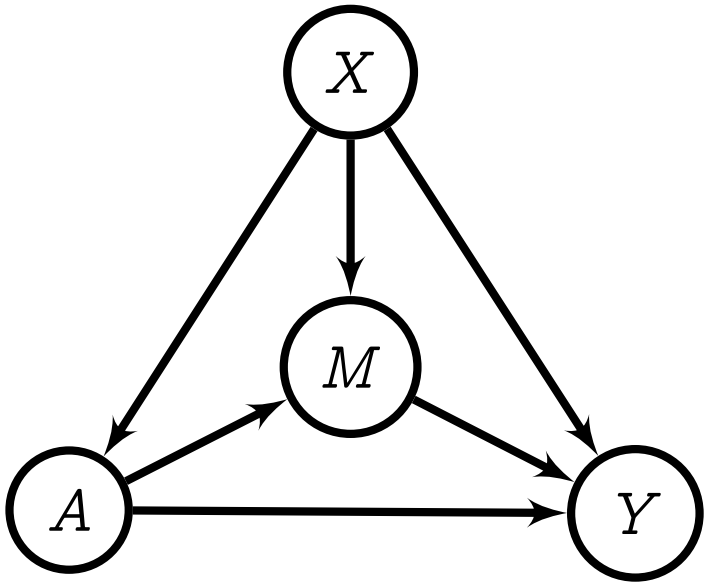


Singh, Xu, G. Sequential kernel embedding for mediated and time-varying dose response curves (Bernoulli 2025)

Mediation analysis: example

US job corps: training for disadvantaged youths:

- X: confounder/context (age, education, marital status, ...)
- A: treatment (training hours)
- Y: outcome (arrests)
- M: mediator (employment)

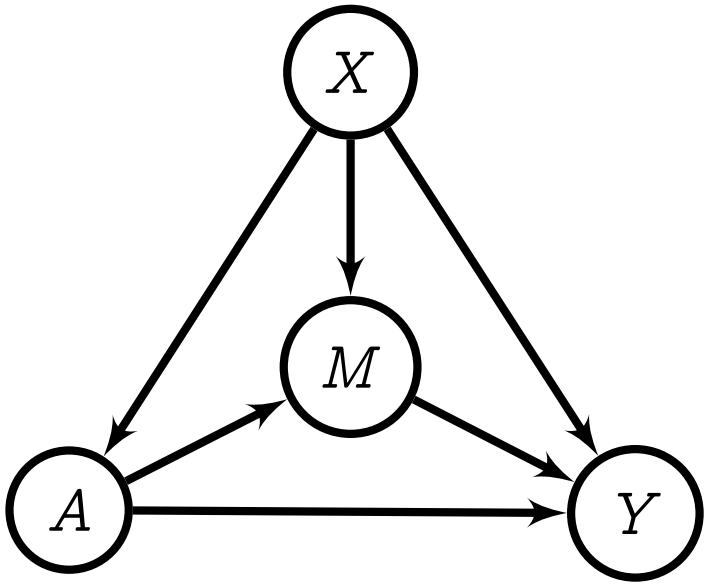


$\gamma_*(a, m, x) \approx \mathbb{E}[Y|A = a, M = m, X = x]$

Mediation analysis: example

US job corps: training for disadvantaged youths:

- X: confounder/context (age, education, marital status, ...)
- A: treatment (training hours)
- Y: outcome (arrests)
- M: mediator (employment)



$$\gamma_*(a, m, x) \approx \mathbb{E}[Y|A = a, M = m, X = x]$$

A quantity of interest, the mediated effect:

$$Y^{\{a', M^{(a)}\}} = \int \gamma_*(a', M, X) \mathrm{d}\mathbb{P}(M|A = a, X)\, d\mathbb{P}(X)$$

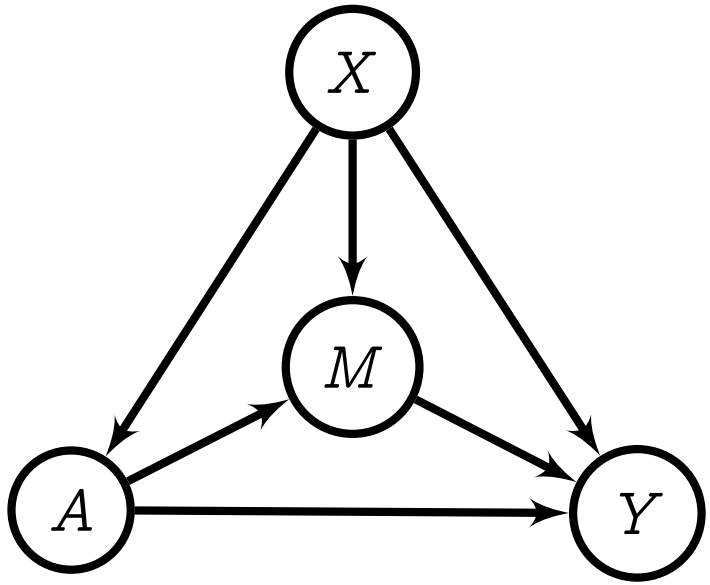
Effect of intervention a', with $M^{(a)}$ as if intervention were a

Singh, Xu, G (2025). Kernel Methods for Multistage Causal Inference: Mediation Analysis and Dynamic Treatment Effects.

Mediation analysis: example

US job corps: training for disadvantaged youths:

- X: confounder/context (age, education, marital status, ...)
- A: treatment (training hours)
- Y: outcome (arrests)
- M: mediator (employment)



$$\gamma_*(a, m, x) \approx \mathbb{E}[Y|A = a, M = m, X = x]$$

A quantity of interest, the mediated effect:

$$Y^{\{a', M^{(a)}\}} = \int \gamma_*(a', M, X) \mathrm{d}\mathbb{P}(M|A = a, X) d\mathbb{P}(X)$$
$$= \langle \gamma_*, \varphi(a') \otimes \mathbb{E}_P\{\mu_{M|A=a,X} \otimes \varphi(X)\}\rangle$$

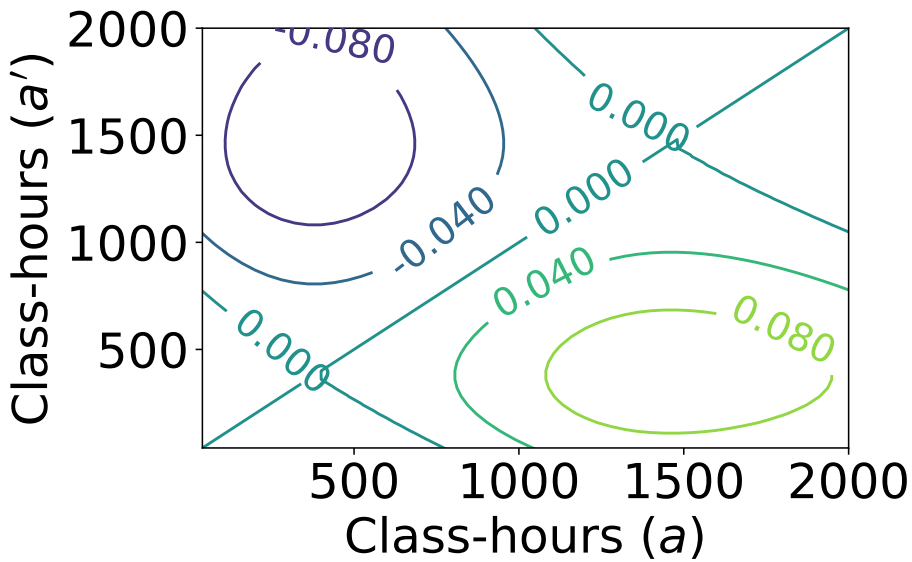
Effect of intervention a', with $M^{(a)}$ as if intervention were a

Singh, Xu, G (2025). Kernel Methods for Multistage Causal Inference: Mediation Analysis and Dynamic Treatment Effects.

Mediation analysis: results

Total effect:

$$\mathrm{TE}(a, a') := \mathbb{E}[Y^{\{a', M^{(a')}\}} - Y^{\{a, M^{(a)}\}}]$$

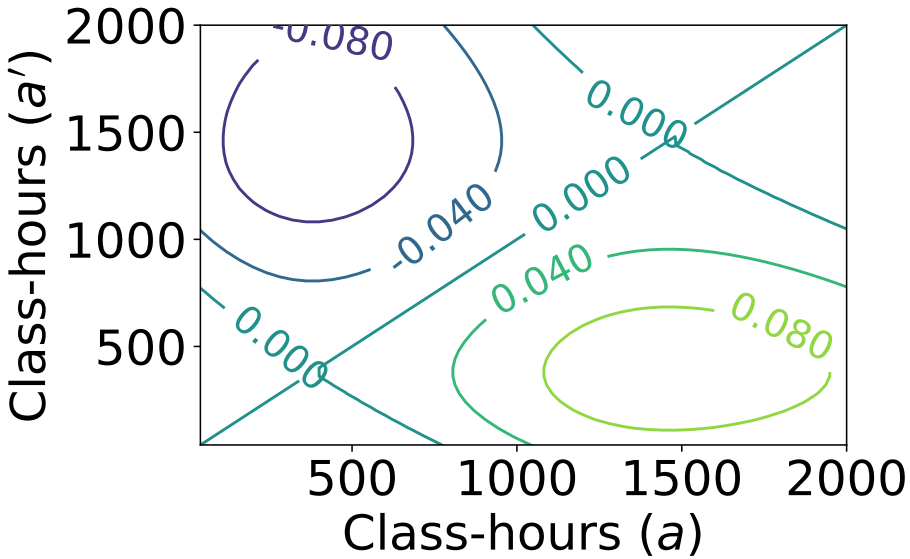


- $a' = 1600$ hours vs $a = 480$ means 0.1 reduction in arrests

Mediation analysis: results

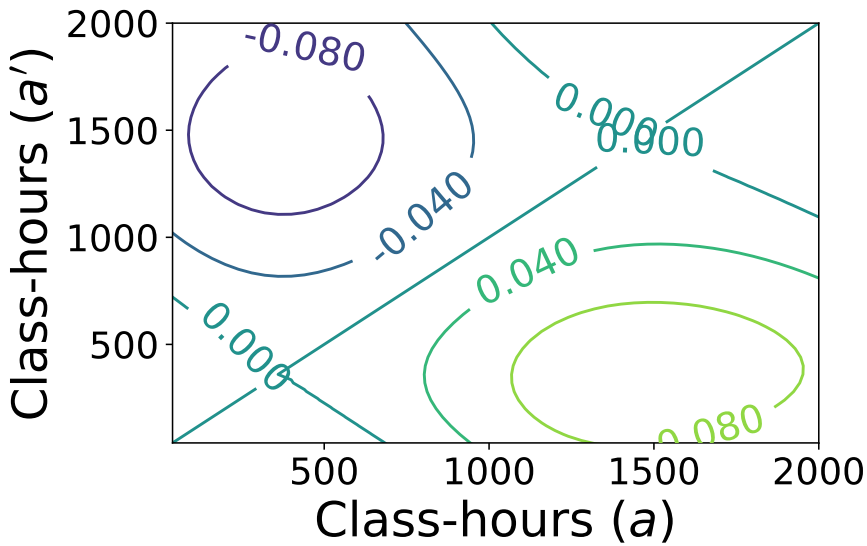
Total effect:

$$\mathrm{TE}(a, a') := \mathbb{E}[\, Y^{\{a', M^{(a')}\}} - Y^{\{a, M^{(a)}\}}]$$



Direct effect:

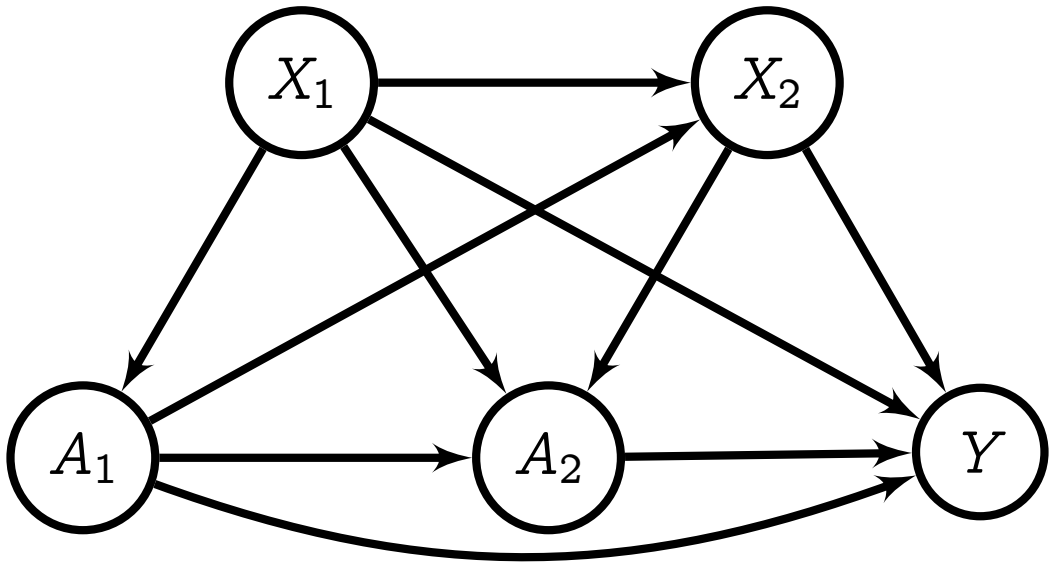
$$\mathrm{DE}(a, a') := \mathbb{E}[\, Y^{\{a', M^{(a)}\}} - Y^{\{a, M^{(a)}\}}]$$



- $a' = 1600$ hours vs $a = 480$ means 0.1 reduction in arrests
- Indirect effect mediated via employment effectively zero

...dynamic treatment effect...

Dynamic treatment effect: sequence A_1, A_2 of treatments.



- potential outcomes $Y^{(a_1)}$, $Y^{(a_2)}$, $Y^{(a_1,a_2)}$,
- counterfactuals $\mathbb{E}\left[Y^{(a_1',a_2')}|A_1 = a_1, A_2 = a_2\right]$...

(c.f. the Robins G-formula)

Instrumental variable regression

Instrumental variable regression

The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2021

© Nobel Prize Outreach. Photo: Paul Kennedy

David Card

Prize share: 1/2

© Nobel Prize Outreach. Photo: Risdon Photography

Joshua D. Angrist

Prize share: 1/4

© Nobel Prize Outreach. Photo: Paul Kennedy

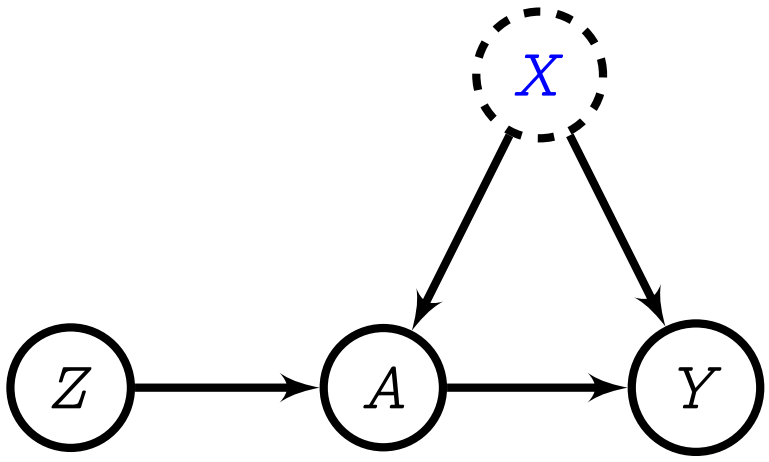
Guido W. Imbens

Prize share: 1/4

The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2021 was divided, one half awarded to David Card "for his empirical contributions to labour economics", the other half jointly to Joshua D. Angrist and Guido W. Imbens "for their methodological contributions to the analysis of causal relationships"

Instrumental variable regression with kernels

- X: unobserved confounder.
- A: treatment
- Y: outcome
- Z: instrument



Assumptions

$$\mathbb{E}[X|Z] = 0$$

$$Z \not\perp\!\!\!\perp A$$

$$(Y \perp\!\!\!\perp Z|A)_{G_{\bar{A}}}$$

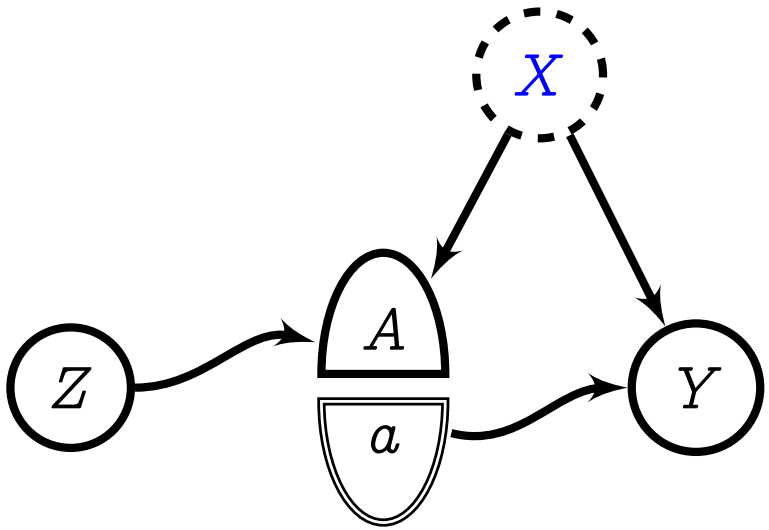
$$Y = \langle \gamma, \phi(A) \rangle + X$$

Instrumental variable regression with kernels

- X: unobserved confounder.
- A: treatment
- Y: outcome
- Z: instrument

Assumptions

$$\mathbb{E}[X|Z] = 0$$
$$Z \not\perp\!\!\!\perp A$$
$$(Y \perp\!\!\!\perp Z|A)_{G_{\bar{A}}}$$
$$Y = \langle \gamma, \phi(A) \rangle + X$$

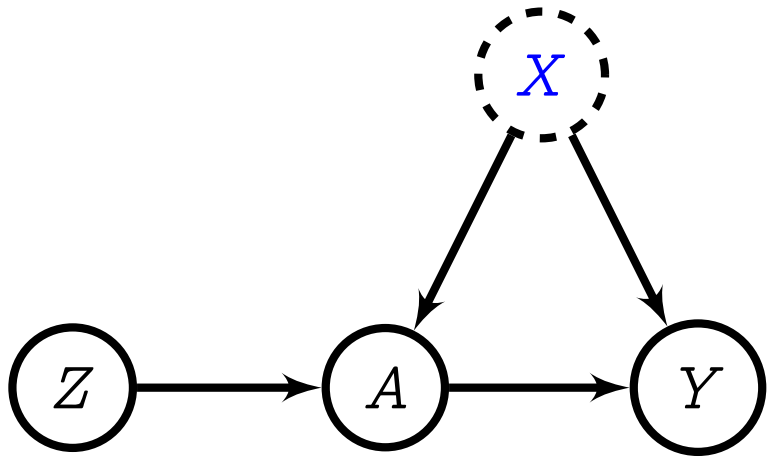


Dose-response curve:

$$\mathrm{DR}(a) = \int \mathbb{E}(Y|X,a)\,dp(X) = \langle \gamma, \phi(a) \rangle$$

Instrumental variable regression with kernels

- X: unobserved confounder.
- A: treatment
- Y: outcome
- Z: instrument



Assumptions

$$\mathbb{E}[X|Z] = 0$$

$$Z \not\perp\!\!\!\perp A$$

$$(Y \perp\!\!\!\perp Z|A)_{G_{\bar{A}}}$$

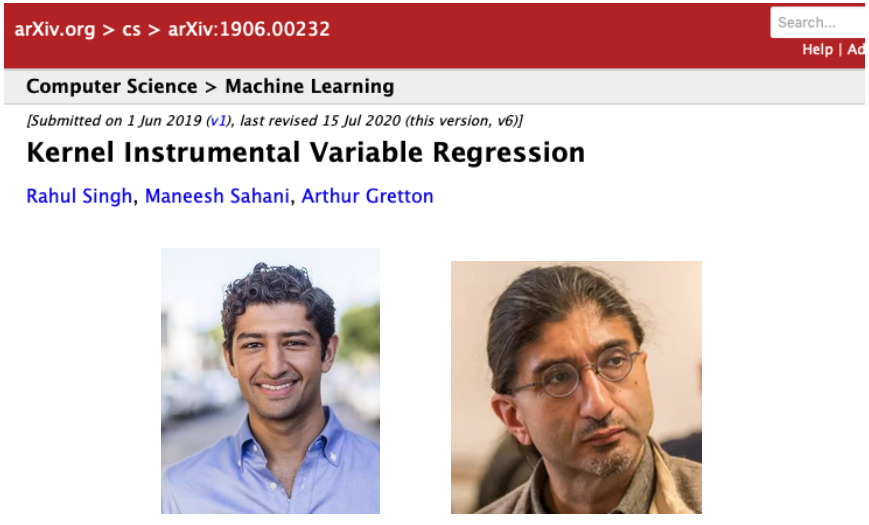
$$Y = \langle \gamma, \phi(A) \rangle + X$$

IV regression: Condition both sides on Z,

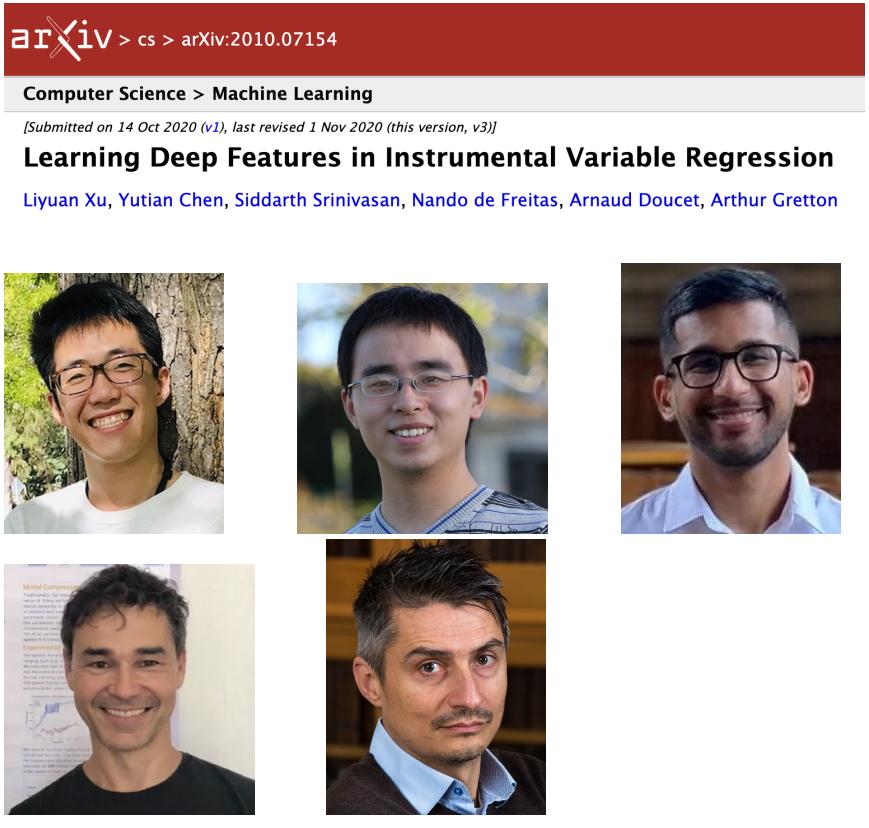
$$\mathbb{E}[Y|Z] = \mathbb{E}\left[\langle \gamma, \varphi(A)]\rangle \,|Z\right] + \underbrace{\mathbb{E}[X|Z]}_{=0}$$

Two-stage least squares for IV regression

Kernel features (NeurIPS 2019):



NN features (ICLR 2021):

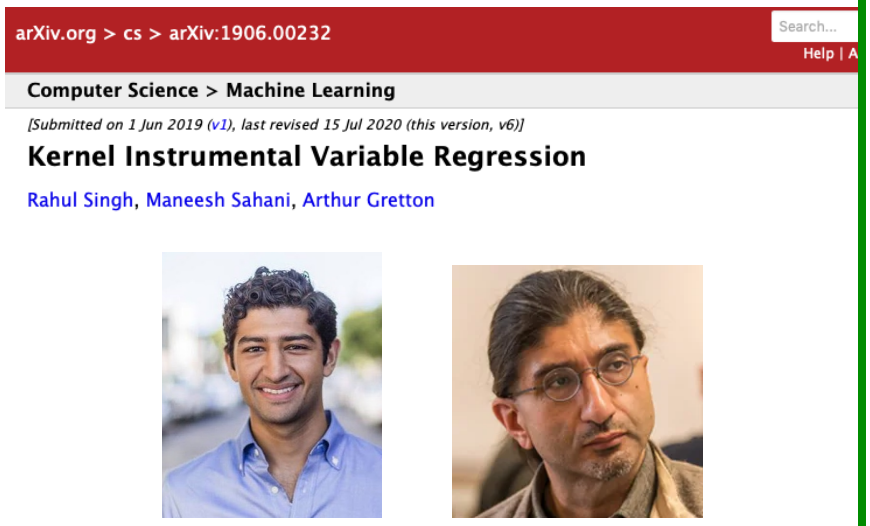


Code for NN and kernel IV methods:

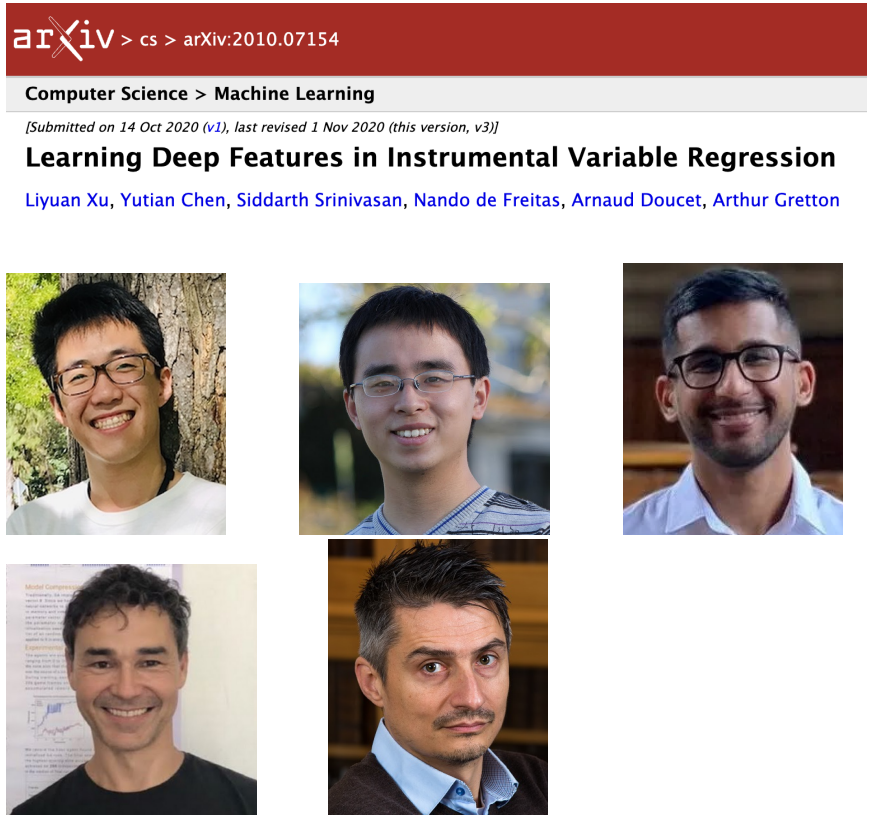
https://github.com/liyuan9988/DeepFeatureIV/

Two-stage least squares for IV regression

Kernel features (NeurIPS 2019):



NN features (ICLR 2021):



Code for NN and kernel IV methods:

https://github.com/liyuan9988/DeepFeatureIV/

IV using kernel features

Stage 2 regression (IV): Solve for γ with RR loss:

$$\hat{\gamma} = \underset{\gamma \in \mathcal{H}_{\mathcal{A}}}{\operatorname{argmin}} \quad \frac{1}{n} \sum_{\ell=1}^{n} \left[\left(y_\ell - \widehat{\mathbb{E}} \left[\langle \gamma, \varphi(A) \rangle_{\mathcal{H}_{\mathcal{A}}} \middle| z_\ell \right] \right)^2 \right] + \lambda_2 \|\gamma\|^2_{\mathcal{H}_{\mathcal{A}}}$$

$$= \underset{\gamma \in \mathcal{H}_{\mathcal{A}}}{\operatorname{argmin}} \quad \frac{1}{n} \sum_{\ell=1}^{n} \left[\left(y_\ell - \left\langle \gamma, \widehat{\mathbb{E}} \left[\varphi(A) | z_\ell \right] \right\rangle_{\mathcal{H}_{\mathcal{A}}} \right)^2 \right] + \lambda_2 \|\gamma\|^2_{\mathcal{H}_{\mathcal{A}}}$$

Xu, Chen, Srinivasan, De Freitas, Doucet, G. (2021) Learning Deep Features in Instrumental Variable Regression

IV using kernel features

Stage 2 regression (IV): Solve for γ with RR loss:

$$\hat{\gamma} = \underset{\gamma \in \mathcal{H}_{\mathcal{A}}}{\operatorname{argmin}} \quad \frac{1}{n} \sum_{\ell=1}^{n} \left[\left(y_\ell - \widehat{\mathbb{E}} \left[\langle \gamma, \varphi(A) \rangle_{\mathcal{H}_{\mathcal{A}}} \middle| z_\ell \right] \right)^2 \right] + \lambda_2 \|\gamma\|^2_{\mathcal{H}_{\mathcal{A}}}$$

$$= \underset{\gamma \in \mathcal{H}_{\mathcal{A}}}{\operatorname{argmin}} \quad \frac{1}{n} \sum_{\ell=1}^{n} \left[\left(y_\ell - \left\langle \gamma, \widehat{\mathbb{E}} \left[\varphi(A) | z_\ell \right] \right\rangle_{\mathcal{H}_{\mathcal{A}}} \right)^2 \right] + \lambda_2 \|\gamma\|^2_{\mathcal{H}_{\mathcal{A}}}$$

Stage 1 regression: define $\widehat{F} \in S_2(\mathcal{H}_{\mathcal{Z}}, \mathcal{H}_{\mathcal{A}})$:

$$\widehat{\mathbb{E}}[\varphi(A)|z] = \widehat{F}\varphi(z)$$

Learn with RR loss:

$$\widehat{F} = \underset{F \in S_2(\mathcal{H}_{\mathcal{Z}}, \mathcal{H}_{\mathcal{A}})}{\operatorname{argmin}} \frac{1}{m} \sum_{\ell=1}^{m} \|\varphi(a_\ell) - F\varphi(z_\ell)\|^2_{\mathcal{H}_{\mathcal{A}}} + \lambda_1 \|F\|^2_{HS}$$

Xu, Chen, Srinivasan, De Freitas, Doucet, G. (2021) Learning Deep Features in Instrumental Variable Regression

IV using kernel features

Stage 2 regression (IV): Solve for γ with RR loss:

$$\hat{\gamma} = \underset{\gamma \in \mathcal{H}_{\mathcal{A}}}{\operatorname{argmin}} \quad \frac{1}{n} \sum_{\ell=1}^{n} \left[\left(y_\ell - \widehat{\mathbb{E}} \left[\langle \gamma, \varphi(A) \rangle_{\mathcal{H}_{\mathcal{A}}} \middle| z_\ell \right] \right)^2 \right] + \lambda_2 \|\gamma\|^2_{\mathcal{H}_{\mathcal{A}}}$$

$$= \underset{\gamma \in \mathcal{H}_{\mathcal{A}}}{\operatorname{argmin}} \quad \frac{1}{n} \sum_{\ell=1}^{n} \left[\left(y_\ell - \left\langle \gamma, \widehat{\mathbb{E}} \left[\varphi(A) | z_\ell \right] \right\rangle_{\mathcal{H}_{\mathcal{A}}} \right)^2 \right] + \lambda_2 \|\gamma\|^2_{\mathcal{H}_{\mathcal{A}}}$$

Stage 1 regression: define $\widehat{F} \in S_2(\mathcal{H}_{\mathcal{Z}}, \mathcal{H}_{\mathcal{A}})$:

$$\widehat{\mathbb{E}}[\varphi(A)|z] = \widehat{F}\varphi(z)$$

Learn with RR loss:

$$\widehat{F} = \underset{F \in S_2(\mathcal{H}_{\mathcal{Z}}, \mathcal{H}_{\mathcal{A}})}{\operatorname{argmin}} \frac{1}{m} \sum_{\ell=1}^{m} \|\varphi(a_\ell) - F\varphi(z_\ell)\|^2_{\mathcal{H}_{\mathcal{A}}} + \lambda_1 \|F\|^2_{HS}$$

Two stage least squares

Xu, Chen, Srinivasan, De Freitas, Doucet, G. (2021) Learning Deep Features in Instrumental Variable Regression

Minimax consistency of kernel IV

Submitted, JMLR (2026):

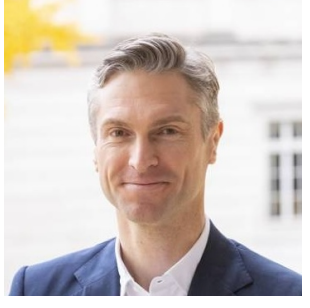
arXiv > stat > arXiv:2411.19653

Statistics > Machine Learning

[Submitted on 29 Nov 2024]

Nonparametric Instrumental Regression via Kernel Methods is Minimax Optimal

Dimitri Meunier, Zhu Li, Tim Christensen, Arthur Gretton



Consistency: assumptions

Define treatment feature covariance:

$$C_A := \mathbb{E}\left[\varphi(A) \otimes \varphi(A)\right] \quad \langle f, C_A g \rangle_{\mathcal{H}_\mathcal{A}} = E_A\left[f(A) g(A)\right]$$

$$C_A = \sum_{i=1}^{\infty} \eta_i \left(\sqrt{\eta_i} e_i\right) \otimes \left(\sqrt{\eta_i} e_i\right)$$

Fine print: $\text{supp}(P_A) = \mathcal{A}$.

Assume problem well specified:

- Denote: $\gamma_* \in \mathcal{H}_\mathcal{A}^c$ where $\mathcal{H}_\mathcal{A}^c \subset \mathcal{H}_\mathcal{A}$, $\quad c \in (1, 2]$

$$\gamma_* = \sum_i \gamma_{*,i} e_i \quad \text{where} \quad \|\gamma_*\|^2_{\mathcal{H}_\mathcal{A}^c} = \sum_{i=1}^{\infty} \frac{\gamma_{*,i}^2}{\eta_i^c} < \infty$$

- Eigenspectrum decay of input feature covariance, $\eta_j \sim j^{-b}$, $b \geq 1$

Consistency: goal

Strong error control:

$$\|\hat{\gamma} - \gamma_*\|_{L_2(A)}$$

Consistency: goal

Strong error control:

$$\|\hat{\gamma} - \gamma_*\|_{L_2(A)}$$

Stage 2 loss naturally expressed in weak metric:

$$\|\mathcal{T}(\hat{\gamma} - \gamma_*)\|_{L_2(Z)}$$

where $\mathcal{T}\gamma = \mathbb{E}(\gamma(A)|Z)$.

Consistency: goal

Strong error control:

$$\|\hat{\gamma} - \gamma_*\|_{L_2(A)}$$

Stage 2 loss naturally expressed in weak metric:

$$\|\mathcal{T}(\hat{\gamma} - \gamma_*)\|_{L_2(Z)}$$

where $\mathcal{T}\gamma = \mathbb{E}(\gamma(A)|Z)$.

By Jensen,

$$\|\mathcal{T}(\hat{\gamma} - \gamma_*)\|_{L_2(Z)} \leq \|\hat{\gamma} - \gamma_*\|_{L_2(A)}$$

Weak control does not imply strong control: need additional link assumption.

Strong/weak geometry, link condition

Strong geometry:

$$\begin{aligned}\|\gamma\|^2_{L_2(A)} &= \mathbb{E}\left[\langle\gamma, \varphi(A)\rangle^2_{\mathcal{H}_\mathcal{A}}\right] \\ &= \langle\gamma, C_A\gamma\rangle_{\mathcal{H}_\mathcal{A}}\end{aligned}$$

where

$$C_A = \mathbb{E}_A\left[\varphi(A) \otimes \varphi(A)\right]$$

Strong/weak geometry, link condition

Strong geometry:

$$\|\gamma\|^2_{L_2(A)} = \mathbb{E}\left[\langle\gamma, \varphi(A)\rangle^2_{\mathcal{H}_\mathcal{A}}\right] = \langle\gamma, C_A\gamma\rangle_{\mathcal{H}_\mathcal{A}}$$

where

$$C_A = \mathbb{E}_A\left[\varphi(A) \otimes \varphi(A)\right]$$

Weak geometry:

$$\|\mathcal{T}\gamma\|^2_{L_2(Z)} = \mathbb{E}\left[\mathbb{E}\left[\langle\gamma, \varphi(A)\rangle_{\mathcal{H}_\mathcal{A}} | Z\right]^2\right] = \langle\gamma, C_M\gamma\rangle_{\mathcal{H}_\mathcal{A}}$$

where

$$C_M = \mathbb{E}\left[\mu_{A|Z} \otimes \mu_{A|Z}\right],$$
$$\mu_{A|Z} := \mathbb{E}\left[\varphi(A)|Z\right].$$

Strong/weak geometry, link condition

Strong geometry:

$$\|\gamma\|^2_{L_2(A)} = \mathbb{E}\left[\langle \gamma, \varphi(A)\rangle^2_{\mathcal{H}_\mathcal{A}}\right] = \langle \gamma, C_A \gamma\rangle_{\mathcal{H}_\mathcal{A}}$$

where

$$C_A = \mathbb{E}_A\left[\varphi(A) \otimes \varphi(A)\right]$$

Weak geometry:

$$\|\mathcal{T}\gamma\|^2_{L_2(Z)} = \mathbb{E}\left[\mathbb{E}\left[\langle \gamma, \varphi(A)\rangle_{\mathcal{H}_\mathcal{A}} \,|Z\right]^2\right] = \langle \gamma, C_M \gamma\rangle_{\mathcal{H}_\mathcal{A}}$$

where

$$C_M = \mathbb{E}\left[\mu_{A|Z} \otimes \mu_{A|Z}\right],$$
$$\mu_{A|Z} := \mathbb{E}\left[\varphi(A)|Z\right].$$

Link condition: $\exists \zeta > 0$, $\theta \geq 1$ such that

$$C_A \preceq \zeta C_M^{1/\theta}$$

Earlier link conditions two-sided $C_A^\gamma \asymp C_M$: Nair, Pereverzev, Tautenhahn (2005); Chen, Reiss (2011).

Convergence result

Model class $\mathcal{M}(c, b, \theta)$: c smoothness of γ_*, b effective dimension, θ ill-posedness.

Convergence result

Model class $\mathcal{M}(c, b, \theta)$: c smoothness of γ_*, b effective dimension, θ ill-posedness.

Assume: Stage 1 samples $m = n^\alpha$ sufficiently large relative to Stage 2 samples n. Then

$$\|\hat{\gamma} - \gamma_*\|^2_{L_2(A)} = O_P(n^{-\frac{c}{c+1/b+\theta-1}}).$$

Convergence result

Model class $\mathcal{M}(c, b, \theta)$: c smoothness of γ_*, b effective dimension, θ ill-posedness.

Assume: Stage 1 samples $m = n^{\alpha}$ sufficiently large relative to Stage 2 samples n. Then

$$\|\hat{\gamma} - \gamma_*\|^2_{L_2(A)} = O_P(n^{-\frac{c}{c+1/b+\theta-1}}).$$

Lower bound (minimax):

$$\inf_{\hat{\gamma}} \sup_{\text{NPIV models in } \mathcal{M}(c,b,\theta)} \|\hat{\gamma} - \gamma_*\|^2_{L_2(A)} \gtrsim n^{-\frac{c}{c+1/b+\theta-1}}$$

For $\theta = 1$ recover nonparametric rate for regression.

IV in reinforcement learning

JMLR (2022)

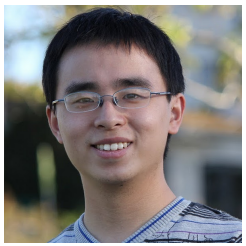
arXiv > cs > arXiv:2105.10148

Computer Science > Machine Learning

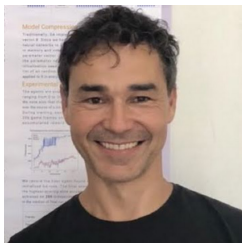
[Submitted on 21 May 2021 (v1), last revised 23 Nov 2022 (this version, v2)]

On Instrumental Variable Regression for Deep Offline Policy Evaluation

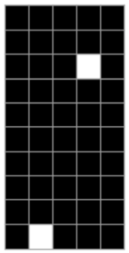
Yutian Chen, Liyuan Xu, Caglar Gulcehre, Tom Le Paine, Arthur Gretton, Nando de Freitas, Arnaud Doucet



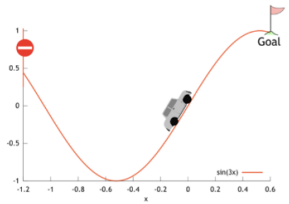




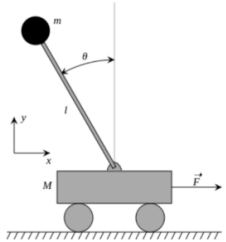

IV in reinforcement learning



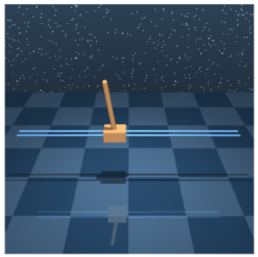
(a) Catch



(b) Mountain Car



(c) Cartpole



(a) Cartpole Swingup

(b) Cheetah Run

(c) Humanoid Run

(d) Walker Walk

Policy evaluation: want Q-value:

$$Q^{\pi}(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t \middle| S_0 = s, A_0 = a\right]$$

for policy $\pi(A|S = s)$.

Osband et al (2019). Behaviour suite for reinforcement learning.https://github.com/deepmind/bsuite

Tassa et al. (2020). dm_control:Software and tasks for continuous control. https://github.com/deepmind/dm_control

Application of IV: reinforcement learning

Q value is a minimizer of Bellman loss

$$\mathcal{L}_{\text{Bellman}} = \mathbb{E}_{SAR}\left[\left(R + \gamma[\mathbb{E}\left[Q^{\pi}(S', A')|S, A\right] - Q^{\pi}(S, A)\right)^2\right].$$

Corresponds to "IV-like" problem

$$\mathcal{L}_{\text{Bellman}} = \mathbb{E}_{YZ}\left[\left(Y - \mathbb{E}[f(X)|Z]\right)^2\right]$$

with

$$\begin{aligned} Y &= R, \\ X &= (S', A', S, A) \\ Z &= (S, A), \\ f_0(X) &= Q^{\pi}(s, a) - \gamma Q^{\pi}(s', a') \end{aligned}$$

RL experiments and data:

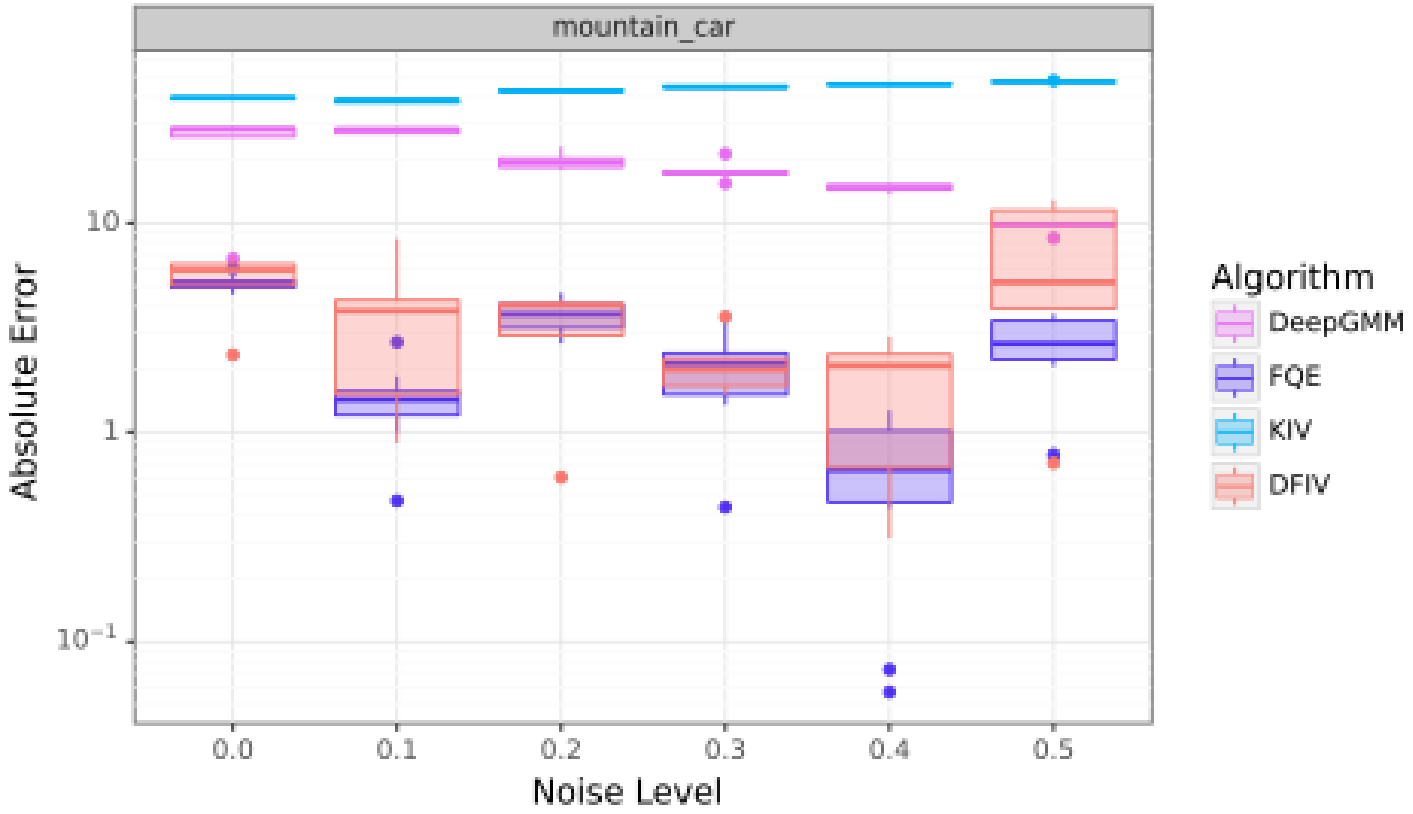
https://github.com/liyuan9988/IVOPEwithACME

Bradtke and Barto (1996). Linear least-squares algorithms for temporal difference learning.

Xu, Chen, Srinivasan, De Freitas, Doucet, G. (2021)

Chen, Xu, Gulcehre, Le Paine, G, De Freitas, Doucet (2022). On Instrumental Variable Regression for Deep Offline Policy Evaluation.

Results on mountain car problem



Good performance compared with FQE.

Warning: IV assumption can fail when regression underfits. See papers for details.

Xu, Chen, Srinivasan, De Freitas, Doucet, G. (2021)

Chen, Xu, Gulcehre, Le Paine, G, De Freitas, Doucet (2022). On Instrumental Variable Regression for Deep Offline Policy Evaluation.

...but seriously, what if there are hidden confounders?

Proxy causal learning

Biometrika (2018), **105**, 4, *pp.* 987–993
Printed in Great Britain

doi: 10.1093/biomet/asy038
Advance Access publication 13 August 2018

Identifying causal effects with proxy variables of an unmeasured confounder

BY WANG MIAO

Guanghua School of Management, Peking University, 5 Summer Palace Road, Haidian District, Beijing 100871, China

mwfy@pku.edu.cn

ZHI GENG

School of Mathematical Sciences, Peking University, 5 Summer Palace Road, Haidian District, Beijing 100871, China

zhigeng@pku.edu.cn

AND ERIC J. TCHETGEN TCHETGEN

Department of Biostatistics, Harvard University, 677 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.

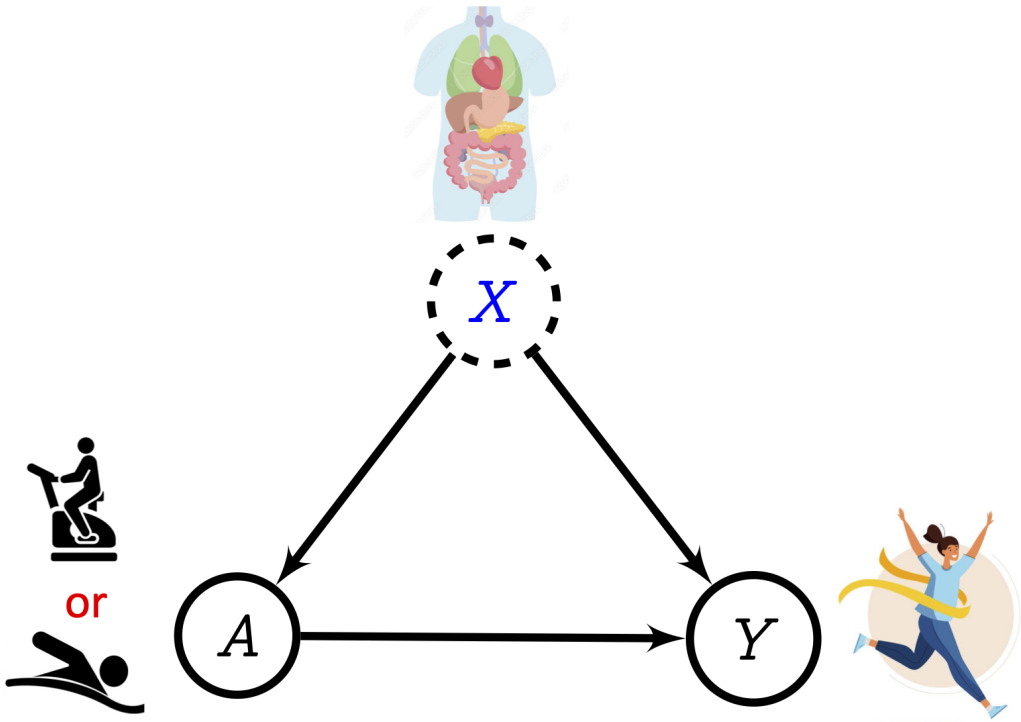
etchetge@hsph.harvard.edu

Proxy Causal Learning: Negative Controls

Unobserved X with (possibly) complex nonlinear effects on A, Y

In this example:

- X: true physical status
- A: exercise regimes
- Y: fitness goal

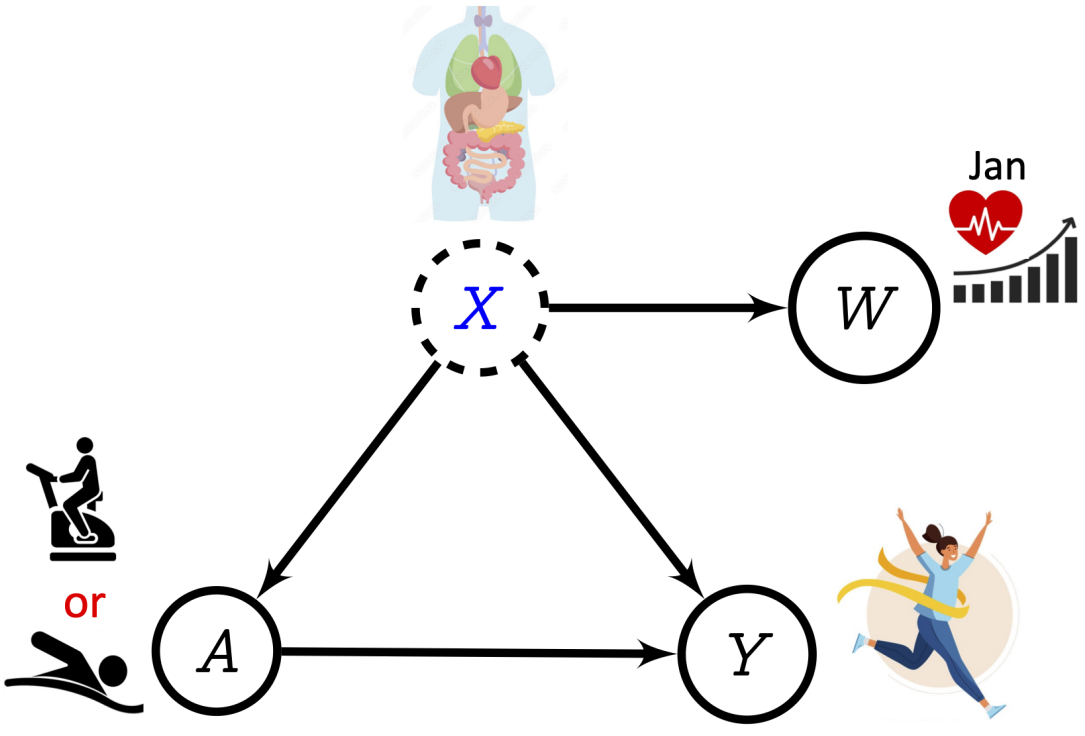


Proxy Causal Learning: Negative Controls

Unobserved X with (possibly) complex nonlinear effects on A, Y

In this example:

- X: true physical status
- A: exercise regimes
- Y: fitness goal
- W: health readings before A

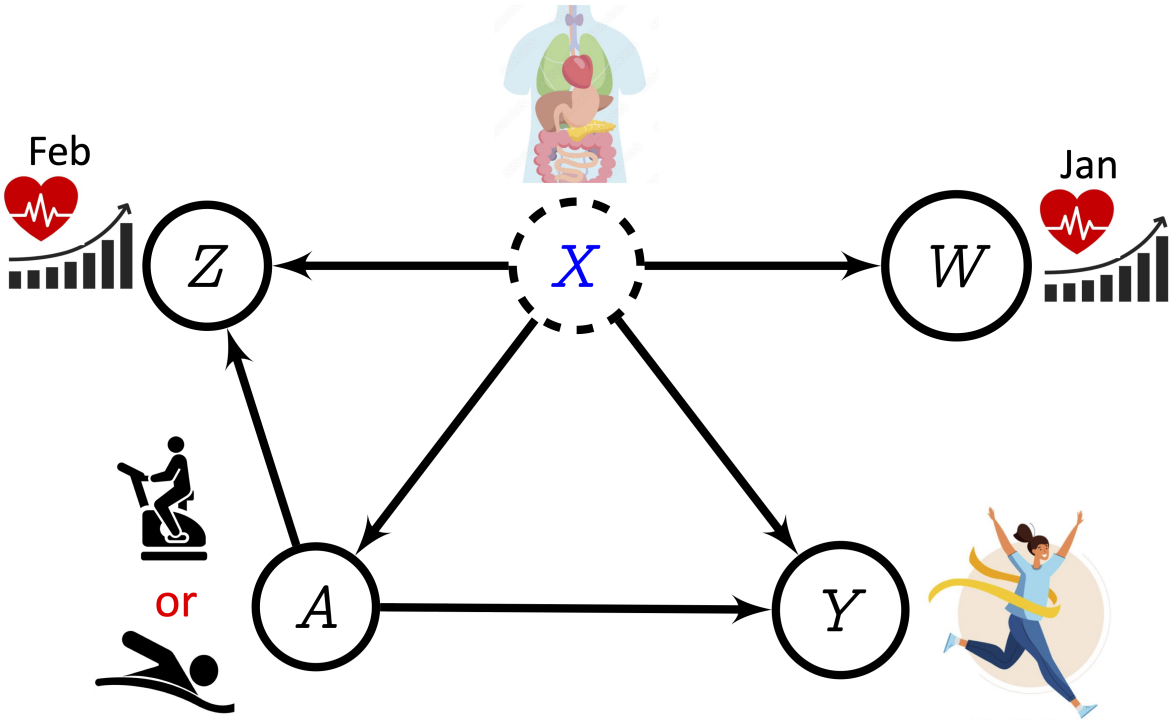


Proxy Causal Learning: Negative Controls

Unobserved X with (possibly) complex nonlinear effects on A, Y

In this example:

- X: true physical status
- A: exercise regimes
- Y: fitness goal
- W: health readings before A
- Z: health readings after A

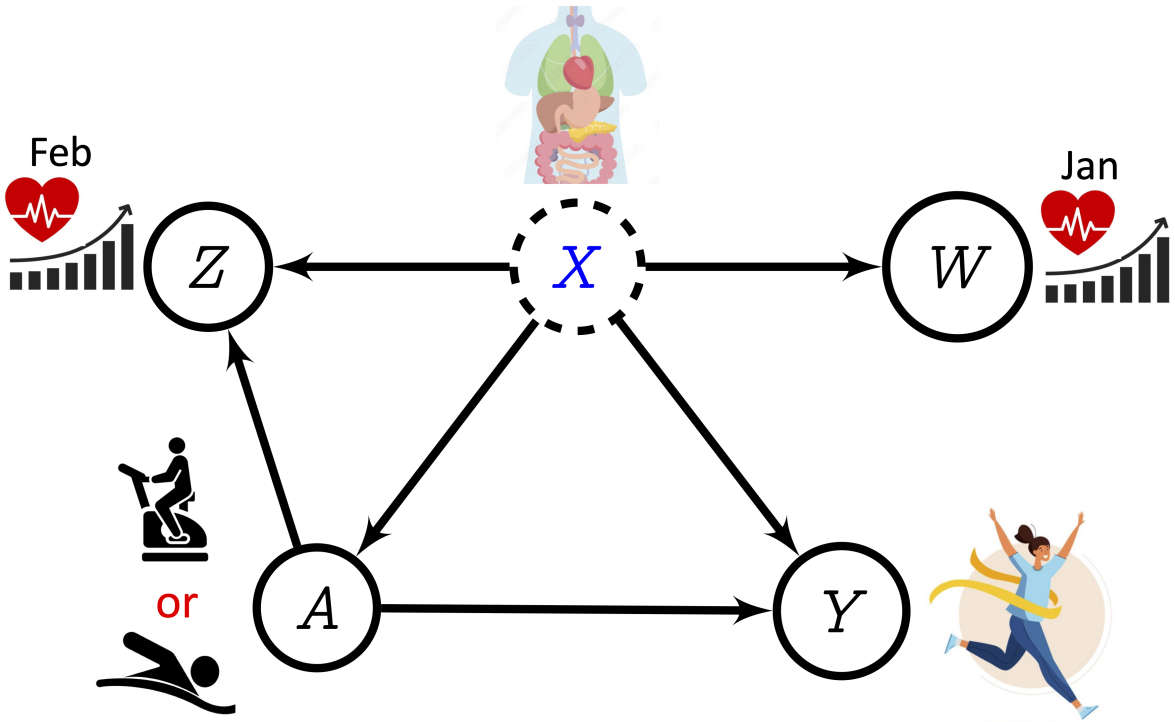


Proxy Causal Learning: Negative Controls

Unobserved X with (possibly) complex nonlinear effects on A, Y

In this example:

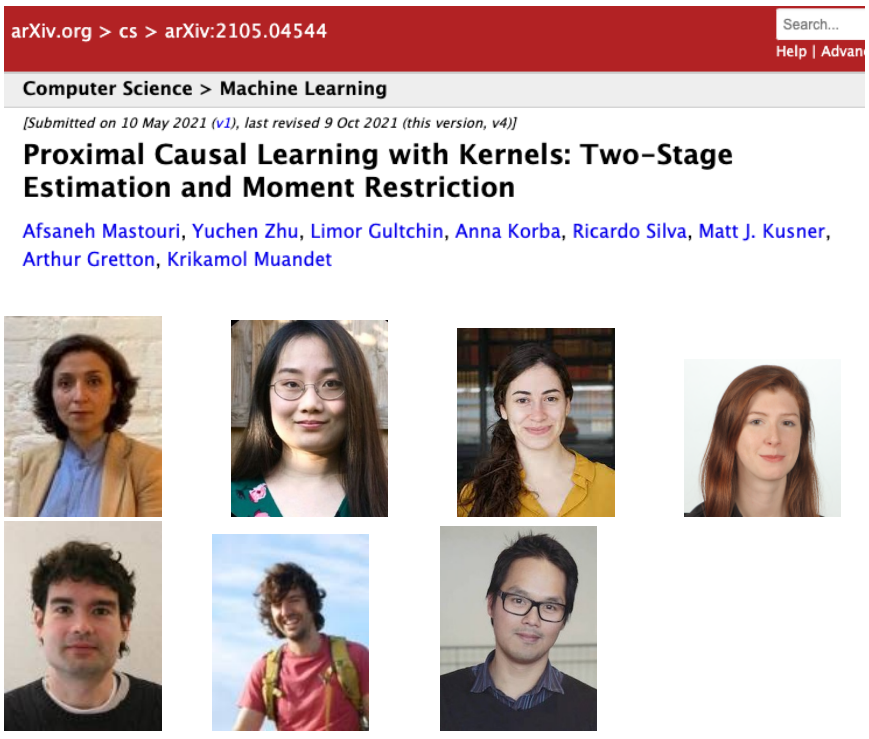
- X: true physical status
- A: exercise regimes
- Y: fitness goal
- W: health readings before A
- Z: health readings after A



$\Longrightarrow$ Can recover $\mathbb{E}(Y^{(a)})$ from observational data

Unobserved confounders: proxy methods

Kernel features (ICML 2021):



NN features (NeurIPS 2021):

Code for NN and kernel proxy methods:

https://github.com/liyuan9988/DeepFeatureProxyVariable/

Conclusions

Kernel (and neural net) solutions:

- ...for dose-response, heterogeneous response, dynamic treatment effects
- ...with treatment A, covariates X, V, multivariate, "complicated"
- Convergence guarantees

Unobserved confounding:

- IV and proxy methods

Code available for all methods

Research support

Work supported by:

The Gatsby Charitable Foundation

Google DeepMind

Google DeepMind

Questions?


THE KERNEL
BREWERY
LONDON
BALTIC PORTER
7.3% ABV
THE KERNEL
EXPORT STOUT